COOPERATIVITY IN NATURAL VERSUS DE NOVO REPEAT PROTEINS AND FUNCTIONAL RAMIFICATIONS

by

Kathryn Geiger-Schuller

A dissertation submitted to the Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy

Baltimore, Maryland

October, 2017

ABSTRACT

Proteins fold and unfold inside living cells. The three-dimension fold of a protein is determined by the order of amino acids in its primary sequence. Folding intermediates are rarified making the cooperativity in protein folding a challenging area of study. Linear repeat proteins make only contacts close in sequence space, thus reducing the number of interacting subunits. One-dimensional Ising models are employed to determine intrinsic and interfacial folding free energies through studying the length-dependence on stability of homopolymeric repeat proteins.

Cooperativity in unnatural helical repeat proteins called *de novo* Helical Repeats (DHRs) is interrogated. These repeat proteins were designed by David Baker's group using the design principles found in the Rosetta software. Both the sequence and structure of DHRs are dissimilar to any observed natural proteins. DHRs fold cooperatively, but they do so in a novel way. Unlike all previously studied naturally-derived consensus repeat proteins, DHRs have favorable intrinsic energies. As a result, DHRs are extremely fast-folding. These results prove that nature could have partitioned stability in a different way, and offer an opportunity for discussion about the evolution of cooperativity and stability in protein folding.

Cooperativity of a naturally-occurring repeat protein, transcription activator-like effectors (TALEs) is also investigated. TALEs bind double stranded DNA one base pair per repeat, and the DNA-binding specificity is determined by two residues in each repeat called repeat-variable diresidues (RVDs).

Consensus TALEs (cTALEs) as well as solubilizing capping motifs are designed. Sequence changes of the RVDs affect the stability and cooperativity of cTALE arrays. cTALEs are moderately cooperative, populating several types of partly folded states.

Population of partly folded cTALE states are tuned for function. Single molecule total internal reflection fluorescence (smTIRF) microscopy and cell-based assays in *S. cerevisiae* show that the cTALEs bind DNA and activate transcription similar to naturally-occurring TALEs. Long movies from smTIRF experiments show binding and unbinding kinetics are multi-phasic suggesting conformational heterogeneity in both free and DNA-bound states. cTALE arrays containing enough repeats to form multiple turns around DNA must unfold to populate higher energy “open” states which are DNA-binding competent. Binding is initiated through a short lived encounter complex which either disassembles or proceeds to a longer lived DNA-bound “locked” state. While this work proves conformational heterogeneity in cTALE binding and unbinding, future work is required to gain insight into the structural details of states involved.

Thesis advisor: Dr. Doug Barrick

Second reader: Dr. Taekjip Ha

Thesis committee: Dr. Scott Bailey

Dr. Rachel Green

Dr. Vincent Hilser

ACKNOWLEDGEMENTS

I would like to thank my thesis advisor, Doug Barrick, for his patient and kind mentorship. He provided me a rich, stimulating graduate school experience, and I cannot fully express my gratitude. I learned to explain complicated things in a simple, clear way from him, and I greatly appreciated his willingness to use many scientific techniques to answer questions.

I would also like to thank the members of my thesis committee Scott Bailey, Rachel Green, Vincent Hilser, and Taekjip Ha, who through meaningful conversations and thoughtful suggestions enhanced the quality of my dissertation research. I would like to thank Dr. Taekjip Ha for the opportunity to collaborate with his lab. Also, I would like to thank my good friend and mentor, Dr. Ananya Majumdar, for NMR training and countless conversations.

I would like to thank all of the members (past and present) of the Barrick Lab. Energetic discussions in lab meetings and journal clubs enriched my time here. I cherish our pedicure sessions and the “One Night Ultimate Werewolf” games we played together. I would also like to thank members of the Ha and Myong lab. Doing some experiments in their space, I appreciated conversations and occasional help finding the TIRF spot. I am grateful for many happy memories working with my friend and colleague, Jaba Mitra. I am also grateful for the support of other students, staff, and faculty in the Program in Molecular Biophysics and the Jenkins Department of Biophysics. I am grateful to Center for Molecular Biophysics, the Integrated Imaging Center, and the Biomolecular NMR Center for training and equipment.

Lastly, I would like to thank my friends and family. Specifically, I am grateful for the unconditional love, support, and occasional sacrifice of my husband, Anthony Schuller. I would also like to thank his parents and family for their love, acceptance, and compassion.

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

ANK	Ankyrin domain
AUC	Analytical Ultracentrifugation
CASP	Critical Assessment of protein Structure Prediction
CD	Circular Dichroism
cTALE	Consensus transcription activator-like effector
DHR	*De novo* Helical Repeat
FRET	Förster resonance energy transfer
GAL4 TAD	GAL4 Transcription activation domain
Gdn HCl	Guanidnium hydrochloride
IDP	Intrinsically Disordered Protein
RVD	Repeat variable diresidue
SM	Single molecule
smTIRF	Single molecule total internal reflection fluorescence
TALE(s)	Transcription activator-like effector(s)
TALENs	Transcription activator-like effector nucleases
TIRF	Total internal reflection fluorescence
TPRs	tetratricopeptide repeat proteins

CHAPTER 1

Introduction

Proteins are biopolymers decorated with 20 unique residues. In a truly incredible chemical reaction, proteins fold into unique structures due to the specific patterning of these 20 side chains. The literature is full of detailed studies outlining the forces responsible for and kinetic barriers to protein folding. While much is understood about protein folding, how proteins achieve cooperativity in protein folding is still a mystery. This work seeks to better understand the energetics of partly folded states and the functional roles such partly folded states play in biology.

1.1 Cell biology mandates proper folding and unfolding

As proteins are synthesized by the ribosome, an unfolded chain protrudes from the exit tunnel. During or after translation, structured domains must fold. There are other times in the protein lifetime when a protein may unfold, or refold (Figure 1.1). Some proteins fold as they are to inserted into membranes while others are unfolded as they are shuttled across membranes[1]. For example, type III secretion machineries unfold substrates during injection into host cells[2]. Some degradation machineries unfold proteins before degradation[3]. Proteins fold and unfold inside living cells[4].

The environment inside cells is clearly different than buffer solutions used in *in vitro* studies. Some proteins require chaperones to fold properly[5,6]. Mechanisms for chaperone-assisted and co-translational folding are active areas

of investigation[6–8]. Although many investigators have predicted the cellular environment to modulate the protein folding energy landscape, several studies comparing folding kinetics and stability measured *in vitro* and *in vivo* find agreement between protein folding inside and out of the cell[9,10].

Against the backdrop of all of this folding, it should be recognized that some proteins never fold. So-called Intrinsically Disordered Proteins (IDPs) do not adopt unique folded structures, but populate ensembles of hetergeneous disordered conformations. Other IDPs only fold as they are binding to another protein[11–13]. While structure is important for function of some biomolecules, IDPs prove that disorder is important for function of other biomolecules. Recent evidence by Rocklin *et al.* suggests some IDPs are actually structured in cells[14]. Examples mentioned by Rocklin *et al.* likely reflect coupled binding and folding of some IDPs bound to interacting partners.

1.2 The protein folding problem: how does sequence determine structure and energy

In his 1972 Nobel Lecture titled “Studies on the Principles that Govern the Folding of Protein Chains”, Christian Anfinsen defines the thermodynamic hypothesis stating “the native conformation is determined by the totality of interatomic interactions and hence by the amino acid sequence, in a given environment”. The sequence of a protein provides instructions to determine the lowest energy structure[15]. In fact, the primary sequence determines all of the possible structures and their populations. If we understood how the sequence of

residues determines the structure of a protein, we could design proteins with new folds or binding scaffolds to target disease-causing proteins.

While the primary sequence determines the three-dimensional fold of a protein, the relationship between sequence and structure is complicated. For example, *E. coli* RNaseH has 88% sequence identity to an ancestral RNaseH[16]. The structural alignment of these proteins (Figure 1.2A) illustrates that proteins with dissimilar sequences can have similar structures. Even a protein with 29% sequence identity to *E. coli* RNaseH (HIV-2 reverse transcriptase) has a similar fold (Figure 1.2B). While the relationship between sequence and structure is complex, computational tools are becoming more successful at structure prediction, as demonstrated by improvements in blind structure prediction benchmarks (Critical Assessment of protein Structure Prediction-CASP)[17]. Some successful strategies include all-atom simulations employing physics-based potentials[18,19], all-atom simulations employing empirically-derived potentials[20], and coarse-grained models[21–23].These advances increase capability to predict the fold of a protein given the primary sequence, but this is only one part of the protein folding problem.

If the goal is to predict the stability of a protein from its primary sequence, an understanding of not only the fully folded and unfolded, but also all partly folded states is required. To gain perspective on the number of possible states, it is useful to consider a model of beads on a string. Figure 1.3A displays 100 beads on a string representing 100 residues in a protein. Making the simplifying assumption that each residue could be in a folded-like conformation or an

unfolded-like conformation, each bead can be either filled in (folded-like) or empty (unfolded-like). Within this framework, Figure 1.3A and 1.3B depict the fully unfolded and folded states of a protein. Even with just two states per residue, there would be 2^{100} possible states, or roughly 10^{30}. This is a huge number of possible states, to gain perspective, it is near the number of bacteria estimated living on planet Earth. The protein folding problem is a major challenge because the goal is to assign energies to all of the one nonillion possible states to predict not only the most likely populated state, but also other states close in energy.

Energy landscapes are a useful tool to visualize the energy of states on the reaction coordinate from fully unfolded to fully folded protein. Figure 1.3D displays a rugged energy landscape where the fully unfolded state is highest in energy and the fully folded state is lowest in energy. In this figure, there are many other low energy states between the unfolded and folded states. Figure 1.3E displays a smooth energy landscape. In this figure, the only low energy states are structurally closer to the folded state than the unfolded state. One might expect that the energy landscape depicted in Figure 1.3E is consistent with natural proteins, however experimental data is commonly consistent with a two-state folding mechanism suggesting proteins have smooth energy landscapes[24]. This means that while there is a vast conformational space available to proteins, many well-studied examples occupy states that look mostly folded or mostly unfolded. This all-or-none behavior is called cooperativity.

1.3 Ising models quantify cooperativity in protein folding.

One simple illustration of cooperativity in protein folding is a chemical denaturation experiment. Figure 1.4 displays a typical equilibrium chemical denaturation of a protein. There are three important regions of this curve. At low denaturant (below 5 M Gdn HCl), the fraction of fully folded protein is near one. At high denaturant (above 6.5 M Gdn HCl), the fraction of fully folded protein is near zero. At intermediate denaturant (between 5.0 and 6.5 M Gdn HCl), the fraction of fully folded protein transitions rapidly from one to zero. In the transition region, proteins populate fully folded and fully unfolded states, but often few or no partly folded conformations. A very rough estimate of the cooperativity in protein folded is given by the slope of this transition, called an m-value. The steeper this slope, the more cooperative the unfolding reaction. Whereas m-values are sensitive to cooperativity, there values are also determined by the change in solvent accessible surface area (SASA) and thus are most closely related to protein size[25].

A more detailed description of cooperativity in protein folding requires determination populations of partly folded states. Often, these populations are too small or these structures are too similar to folded or unfolded states to be measured. Figure 1.5 depicts partly folded microstates of a protein with three units of structure (N, R, and C). In many cases, only the fully folded microstate and fully unfolded microstate are populated enough to measure.

Although populations of rare microstates cannot be measured directly, they can be indirectly quantified using fitted parameters from Ising analysis. In a one-dimensional Ising model, each subunit can exist in the folded or unfolded state.

For a protein with *n* subunits, the total number of microstates is 2^n. The energy of each microstate is determined by equilibrium constants related to the intrinsic folding energy of each subunit (ΔG_N, ΔG_R, and ΔG_C in Figure 1.5) as well as the coupling energy between adjacent subunits ($\Delta G_{i, i+1}$ in Figure 1.5). When intrinsic and interfacial energies are known, it is possible to calculate the energies of all of the microstates, no matter how rare they are. With this formalism, one can access a quantitative description of cooperativity. By studying the structural determinants of the Ising parameters (i.e., the intrinsic and interfacial energies), we can learn about the molecular determinants of cooperativity.

1.4 Repeat proteins are a simplified system useful for studies of cooperativity.

One way to simplify studying cooperativity in protein folding is to reduce the number of subunits with which any single subunit is able to interact. Linear repeat proteins make only contacts that are close in sequence space, thus limiting interacting subunits to only the previous and subsequent repeats. Repeats can be added or removed without disturbing the overall fold of the protein[26]. Although natural repeat proteins have similar folds from repeat to repeat, the sequence identity from repeat to repeat can be low. In these natural repeat protein systems with heterogeneity, the intrinsic energy of each repeat as well as interfacial energies of adjacent repeats are likely to be different for each unique repeat and interface. Intrinsic and interfacial energies are difficult to determine in these heterogeneous systems[27].

To further simplify repeat protein systems, homogenous consensus repeat proteins have been designed. In consensus repeat proteins, the sequence of every repeat is identical. Consensus repeat proteins are designed by generating alignments using a large number of sequences from a given repeat family. Consensus repeat proteins have been designed for several types of repeat proteins including Ankyrin[28,29], TPR[30,31], and LRR (Thuy Dao, unpublished data). In many cases, arrays require solubilizing N- and C-terminal capping repeats to prevent consensus array association and higher-order aggregation [29,30,32–35].

By measuring the length- and capping-dependence on consensus repeat protein stability, intrinsic and interfacial energies can be determined using the Ising formalism described above (Figure 1.6)[36]. Several published reports provide intrinsic and interfacial energies for ankyrin repeat[29] and TPR[30,32] protein families. In all of these studies, intrinsic folding free energies are unfavorable and are offset by favorable interfacial free energies between repeats. This means that single repeats do not fold. An array containing many repeats can fold, and the energetic driving force for folding is provided by favorable coupling energies between adjacent repeats.

It is clear that folding cooperativity is a requirement of natural proteins. Studies of consensus proteins designed using natural sequence information provide insight into how nature solved the problem of cooperativity. By studying designed proteins that share no resemblance to natural proteins, one can ask if joining unstable units folding via high coupling is the only way, or best way, to achieve high cooperativity.

1.5 Functional instability and partly folded states in action

It is clear that promoting too many partly folded states leads to protein aggregation and disease[4]. Nature selected for cooperative proteins, but the level of cooperativity may be finely tuned for function. Functional instability arises in cases where high-energy partly folded states are active conformations.

1.6 Overview

Understanding how folding energy partitions into intrinsic and interfacial energies is key for building a quantitative picture of cooperativity in protein folding. The work in this thesis describes how changes in primary sequence affect distributions of partly folded states in *de novo* designed repeat proteins as well as naturally-derived consensus repeat proteins. A central focus is understanding the functional relevance of partly folded states. Techniques used include circular dichroism (CD) spectroscopy, fluorescence stopped-flow spectroscopy, single molecule total internal reflection (smTIRF) spectroscopy, deterministic simulations, and cell-based assays. Not only does the work in this thesis quantify cooperativity in two unique repeat protein systems, but it also provides insight into how cooperativity is finely tuned for function.

In Chapter 2, I determine intrinsic and interfacial free for unnatural helical repeat proteins called *De novo* Helical Repeats (DHRs). I find that DHRs fold cooperatively, but they do so in a novel way. In chapter 3, I determine intrinsic and interfacial free energies for a DNA-binding repeat protein family, transcription activator-like effectors (TALEs), demonstrating that TALEs are moderately cooperative, and that changes to specificity-conferring residues affect the stability

and cooperativity of TALE arrays in an unexpected way. In chapter 4, I describe conformational heterogeneity in free and DNA-bound TALEs, using single-molecule TIRF, analyze the data using deterministic kinetic modeling, and interpret the results in terms of the structure and energetics of these unique proteins.

1.7 References

1. De Geyter, J. *et al.* Protein folding in the cell envelope of Escherichia coli. *Nat. Microbiol.* **1,** 16107 (2016).
2. Deng, W. *et al.* Assembly, structure, function and regulation of type III secretion systems. *Nat. Rev. Microbiol.* **15,** 323–337 (2017).
3. Olivares, A. O., Baker, T. A. & Sauer, R. T. Mechanistic insights into bacterial AAA+ proteases and protein-remodelling machines. *Nat. Rev. Microbiol.* **14,** 33–44 (2016).
4. Gruebele, M., Dave, K. & Sukenik, S. Globular Protein Folding In Vitro and In Vivo. *Annu. Rev. Biophys.* **45,** 233–251 (2016).
5. Balchin, D., Hayer-Hartl, M. & Hartl, F. U. In vivo aspects of protein folding and quality control. *Science* **353,** aac4354 (2016).
6. Horowitz, S., Koldewey, P., Stull, F. & Bardwell, J. C. Folding while bound to chaperones. *Curr. Opin. Struct. Biol.* **48,** 1–5 (2017).
7. Nilsson, O. B. *et al.* Cotranslational folding of spectrin domains via partially structured states. *Nat. Struct. Mol. Biol.* **24,** 221–225 (2017).
8. Clark, P. L. & Elcock, A. H. Molecular chaperones: providing a safe place to weather a midlife protein-folding crisis. *Nat. Struct. Mol. Biol.* **23,** 621–623 (2016).
9. Guo, M., Xu, Y. & Gruebele, M. Temperature dependence of protein folding kinetics in living cells. *Proc. Natl. Acad. Sci. U. S. A.* **109,** 17863–17867 (2012).

10. Danielsson, J. & Oliveberg, M. Comparing protein behaviour in vitro and in vivo, what does the data really tell us? *Curr. Opin. Struct. Biol.* **42,** 129–135 (2017).
11. Shammas, S. L., Crabtree, M. D., Dahal, L., Wicky, B. I. M. & Clarke, J. Insights into Coupled Folding and Binding Mechanisms from Kinetic Studies. *J. Biol. Chem.* **291,** 6689–6695 (2016).
12. Smock, R. G. & Gierasch, L. M. Sending signals dynamically. *Science* **324,** 198–203 (2009).
13. Tantos, A., Han, K.-H. & Tompa, P. Intrinsic disorder in cell signaling and gene transcription. *Mol. Cell. Endocrinol.* **348,** 457–465 (2012).
14. Rocklin, G. J. *et al.* Global analysis of protein folding using massively parallel design, synthesis, and testing. *Science* **357,** 168–175 (2017).
15. Haber, E. & Anfinsen, C. B. Side-chain interactions governing the pairing of half-cystine residues in ribonuclease. *J. Biol. Chem.* **237,** 1839–1844 (1962).
16. Hart, K. M. *et al.* Thermodynamic system drift in protein evolution. *PLoS Biol.* **12,** e1001994 (2014).
17. Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T. & Tramontano, A. Critical assessment of methods of protein structure prediction: Progress and new directions in round XI. *Proteins* **84 Suppl 1,** 4–14 (2016).
18. Lindorff-Larsen, K., Piana, S., Dror, R. O. & Shaw, D. E. How fast-folding proteins fold. *Science* **334,** 517–520 (2011).
19. Bowman, G. R. & Pande, V. S. Protein folded states are kinetic hubs. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 10890–10895 (2010).

20. Das, R. & Baker, D. Macromolecular modeling with rosetta. *Annu. Rev. Biochem.* **77,** 363–382 (2008).
21. Friedrichs, M. S. & Wolynes, P. G. Toward protein tertiary structure recognition by means of associative memory hamiltonians. *Science* **246,** 371–373 (1989).
22. Schafer, N. P., Kim, B. L., Zheng, W. & Wolynes, P. G. Learning To Fold Proteins Using Energy Landscape Theory. *Isr. J. Chem.* **54,** 1311–1337 (2014).
23. Clementi, C., Nymeyer, H. & Onuchic, J. N. Topological and energetic factors: what determines the structural details of the transition state ensemble and 'en-route' intermediates for protein folding? An investigation for small globular proteins. *J. Mol. Biol.* **298,** 937–953 (2000).
24. Sosnick, T. R. & Barrick, D. The folding of single domain proteins--have we reached a consensus? *Curr. Opin. Struct. Biol.* **21,** 12–24 (2011).
25. Myers, J. K., Pace, C. N. & Scholtz, J. M. Denaturant m values and heat capacity changes: relation to changes in accessible surface areas of protein unfolding. *Protein Sci. Publ. Protein Soc.* **4,** 2138–2148 (1995).
26. Tripp, K. W. & Barrick, D. The tolerance of a modular protein to duplication and deletion of internal repeats. *J. Mol. Biol.* **344,** 169–178 (2004).
27. Mello, C. C. & Barrick, D. An experimentally determined protein folding energy landscape. *Proc. Natl. Acad. Sci. U. S. A.* **101,** 14102–14107 (2004).

28. Tripp, K. W. & Barrick, D. Rerouting the folding pathway of the Notch ankyrin domain by reshaping the energy landscape. *J. Am. Chem. Soc.* **130,** 5681–5688 (2008).
29. Aksel, T., Majumdar, A. & Barrick, D. The contribution of entropy, enthalpy, and hydrophobic desolvation to cooperativity in repeat-protein folding. *Struct. Lond. Engl. 1993* **19,** 349–360 (2011).
30. Marold, J. D., Kavran, J. M., Bowman, G. D. & Barrick, D. A Naturally Occurring Repeat Protein with High Internal Sequence Identity Defines a New Class of TPR-like Proteins. *Struct. Lond. Engl. 1993* (2015). doi:10.1016/j.str.2015.07.022
31. Main, E. R. G., Xiong, Y., Cocco, M. J., D'Andrea, L. & Regan, L. Design of stable alpha-helical arrays from an idealized TPR motif. *Struct. Lond. Engl. 1993* **11,** 497–508 (2003).
32. Kajander, T., Cortajarena, A. L., Main, E. R. G., Mochrie, S. G. J. & Regan, L. A new folding paradigm for repeat proteins. *J. Am. Chem. Soc.* **127,** 10188–10190 (2005).
33. Wetzel, S. K., Settanni, G., Kenig, M., Binz, H. K. & Plückthun, A. Folding and unfolding mechanism of highly stable full-consensus ankyrin repeat proteins. *J. Mol. Biol.* **376,** 241–257 (2008).
34. Mosavi, L. K. & Peng, Z.-Y. Structure-based substitutions for increased solubility of a designed protein. *Protein Eng.* **16,** 739–745 (2003).

35. Tripp, K. W. & Barrick, D. Enhancing the stability and folding rate of a repeat protein through the addition of consensus repeats. *J. Mol. Biol.* **365,** 1187–1200 (2007).
36. Aksel, T. & Barrick, D. Analysis of repeat-protein folding using nearest-neighbor statistical mechanical models. *Methods Enzymol.* **455,** 95–125 (2009).
37. Ramos-Morales, F. Impact of Salmonella enterica Type III Secretion System Effectors on the Eukaryotic Host Cell. *International Scholarly Research Notices* (2012). Available at: https://www.hindawi.com/journals/isrn/2012/787934/. (Accessed: 15th August 2017)
38. Katayanagi, K. *et al.* Structural details of ribonuclease H from Escherichia coli as refined to an atomic resolution. *J. Mol. Biol.* **223,** 1029–1052 (1992).
39. Ren, J. *et al.* Structure of HIV-2 reverse transcriptase at 2.35-A resolution and the mechanism of resistance to non-nucleoside inhibitors. *Proc. Natl. Acad. Sci. U. S. A.* **99,** 14410–14415 (2002).

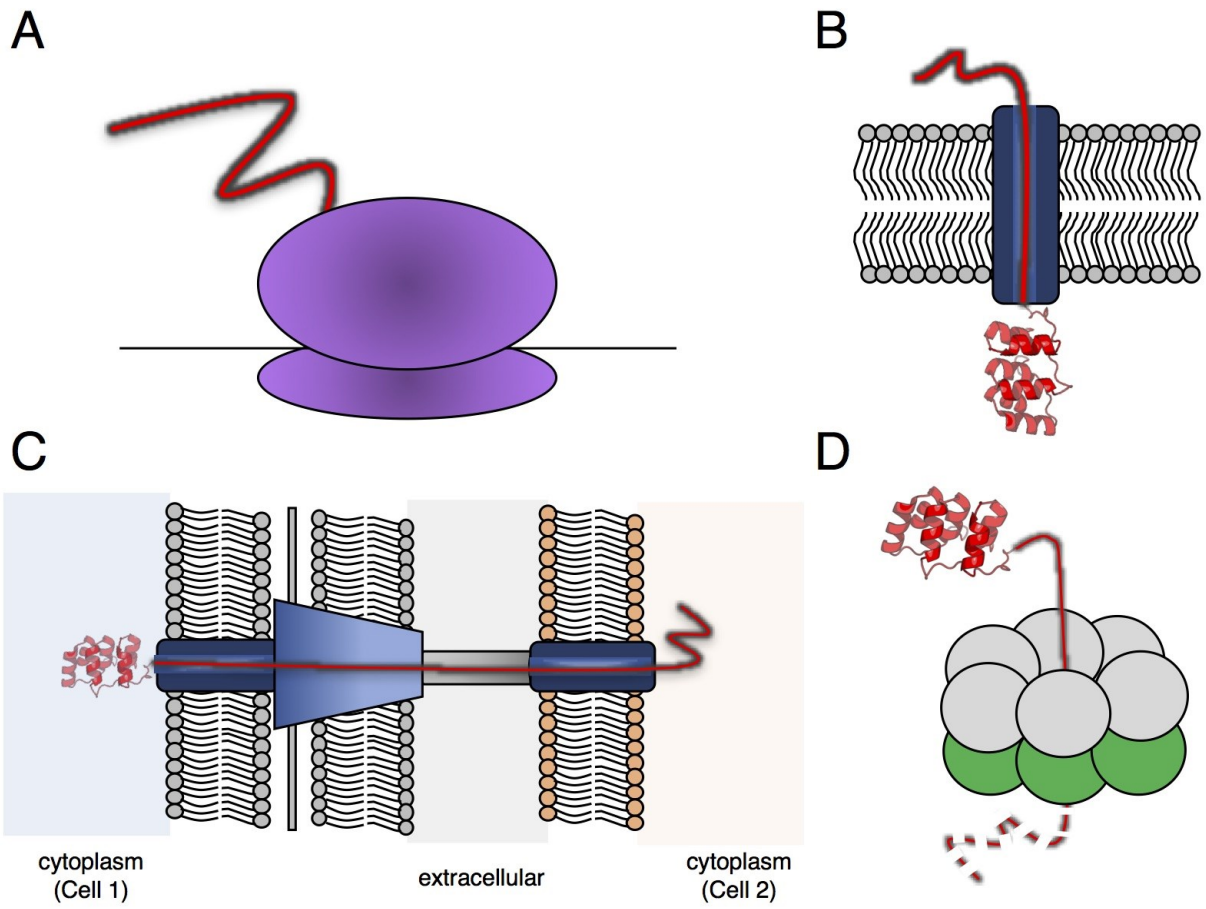


Figure 1.1 Proteins fold and unfold inside living cells. Cartoons describing cellular events that require folding, unfolding, and refolding. (A) Proteins are translated unfolded, and proteins which require structure for function must fold. (B) Proteins translocated across the lipid bilayer of the ER via the translocon must be unfolded. (C) Type IIIS Secretion machinery unfolds to inject proteins into host cells. One specific example is a *Salmonella* cell injecting a toxic protein into a human intestinal host cell[37]. (D) ClpX protease unfolds proteins as it degrades them[3].

Figure 1.2 Proteins with high and low sequence identity can adopt similar folds. (A) *E. coli* RNaseH (PDB: 2RN2)[38] has 88% sequence identity to an ancestral version of RNaseH (PDB: 4LY7)[16]. While there is variation in the primary sequence, the structural alignment of these proteins shows high structural similarity. (B) *E. coli* RNaseH has 29% sequence identity to HIV-2 Reverse Transcriptase (PDB: 1MU2)[39]. Even though there is very low sequence identity in the primary sequence, the structural alignment of these proteins shows high structural similarity.

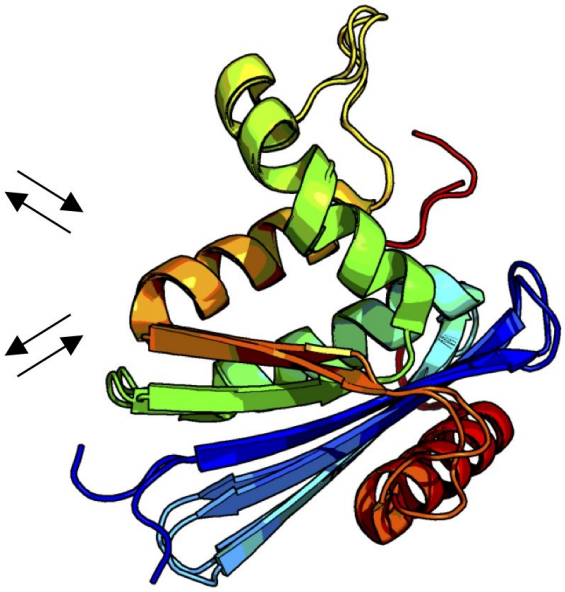
A

***E. coli* RnaseH**

MLKQVEIFTDGSCLGNPGPGGYGAIL
RYRGREKTFSAGYTRTTNNRMELMAA
IVALEALKEHCEVILSTDSQYVRQGI
TQWIHNWKKRGWKTADKKPVKNVDLW
QRLDAALGQHQIKWEWVKGHAGHPEN
ERCDELARAAAMNPTLEDTGYQVEV

Ancestral RnaseH

MLKQVEIFTDGSCLGNPGPGGYGAIL
RYKQHEKTFSAGFHLTTNNRMELMAA
IVALESLKQPCEVILSTDSQYVRQGI
TQWIHNWKKRGWKTADKKPVKNVDLW
QRLDAAIQRHTINWKWVKGHAGHPEN
ERCDELARTAAESPTLEDVGYQPNA



B

***E. coli* RnaseH**

MLKQVEIFTDGSCLGNPGPGGYGAIL
RYRGREKTFSAGYTRTTNNRMELMAA
IVALEALKEHCEVILSTDSQYVRQGI
TQWIHNWKKRGWKTADKKPVKNVDLW
QRLDAALGQHQIKWEWVKGHAGHPEN
ERCDELARAAAMNPTLEDTGYQVEV

HIV-2 Reverse Transcriptase

GDPIPGAETFYTDGSCNRQSKEGKAG
YVTDRGKDKVKKLEQTTNQQAELEAF
AMALTDSGPKVNIIVDSQYVMGIVAS
QPTESESKIVNQIIEEMIKKEAIYVA
WVPAHKGIGGNQEVDHLVSQGI

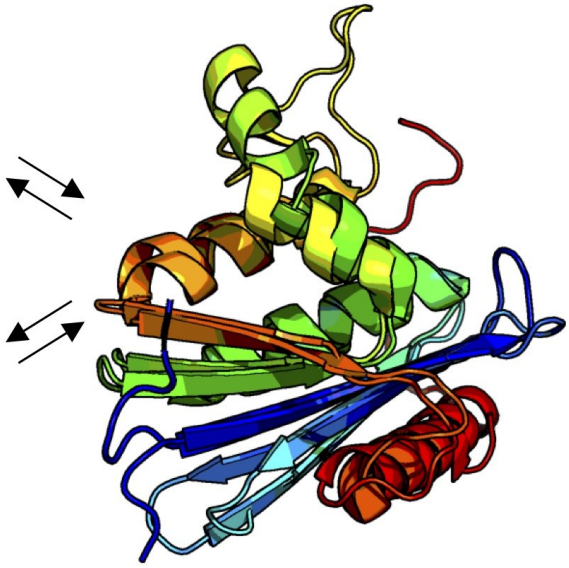
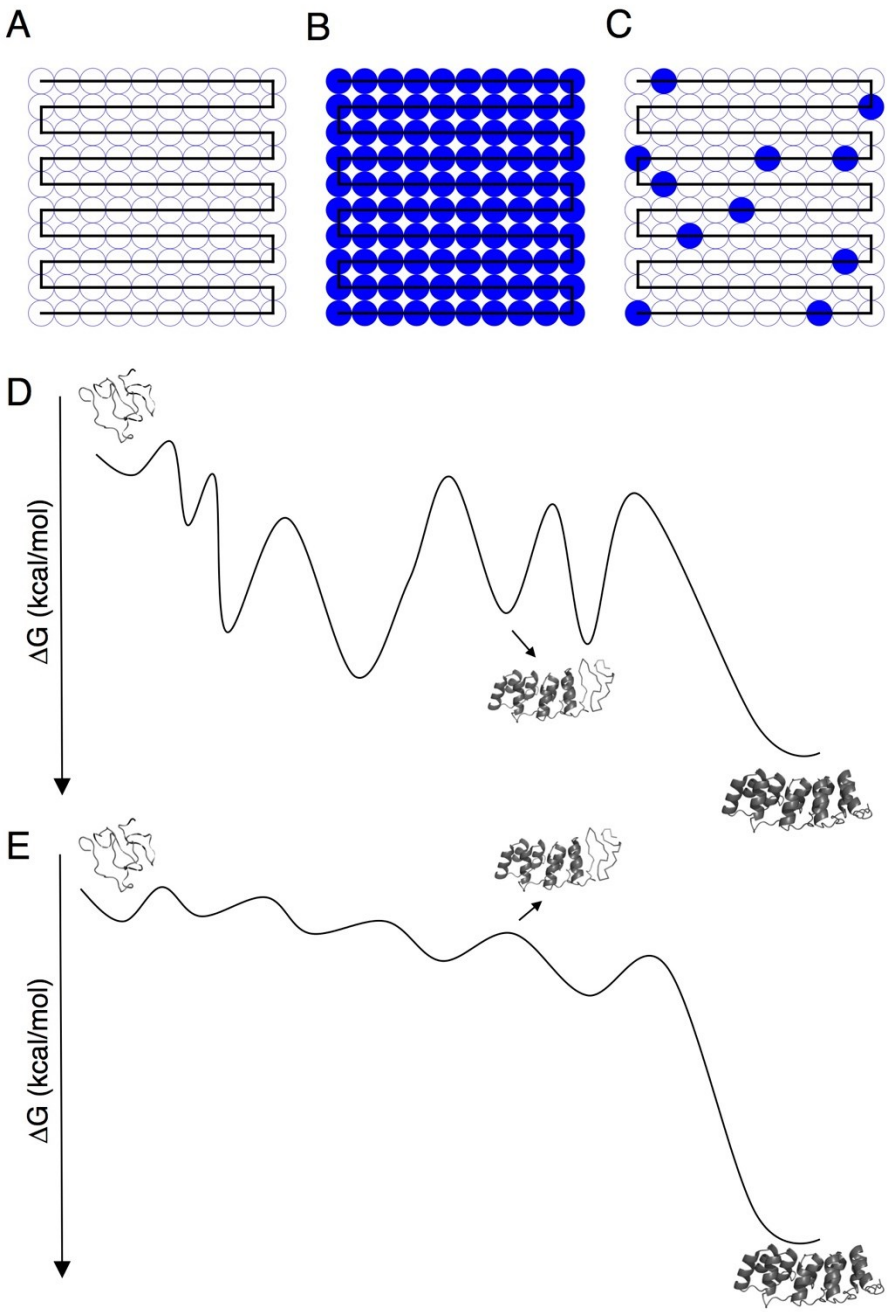


Figure 1.3 Beads on a string model and energy landscapes. (A-C) Beads on a string model depicting fully unfolded state (A), fully folded state (B), and one partly folded state (C). (D-E) Possible protein folding energy landscapes where the x-axis is a folding reaction coordinate and the y-axis is free energy. (D) Rugged energy landscape with many minima spread throughout the folding reaction coordinate. (E) Smooth energy landscape with smaller energy variation where the energies of minima are similar to the fully folded state.


A
B
C
D
ΔG (kcal/mol)
E
ΔG (kcal/mol)

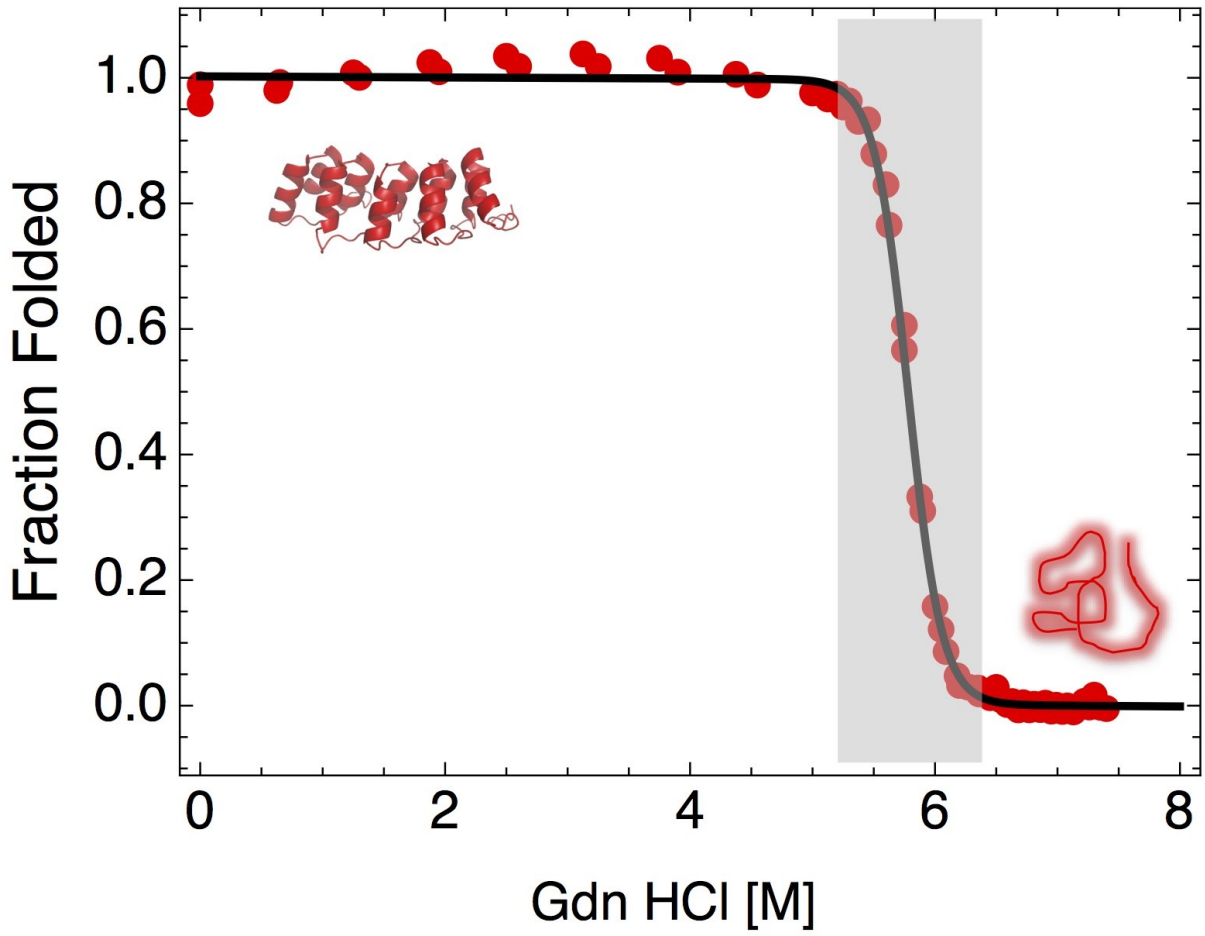


Figure 1.4 Chemical denaturation of a protein. Protein unfolding transition in response to the chemical denaturant GdnHCl. Three regions of this unfolding curve are relevant: first, the region at low denaturant (below 5.0 M GdnHCl in this example) where the fraction folded is roughly one; second, the region above 6.5 M GdnHCl where fraction folded is roughly zero; third, the region between 5.0 and 6.5 M Gdn HCl where fraction folded transitions rapidly from one to zero. (Data collected using protein DHR54 NR_2C to be discussed later in Chapter 2.)

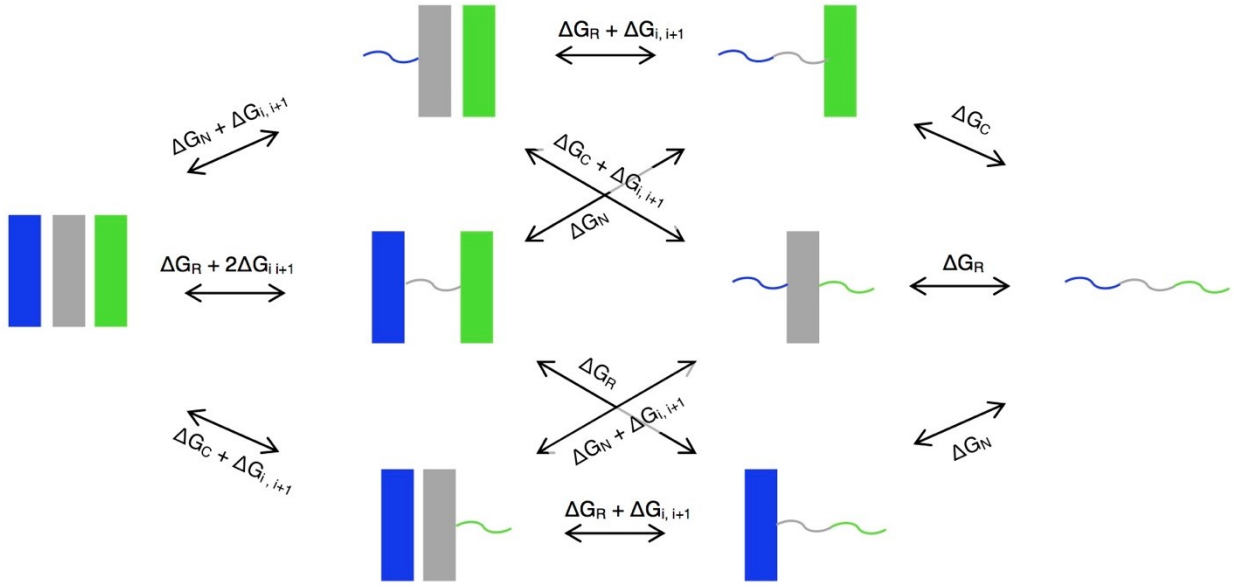
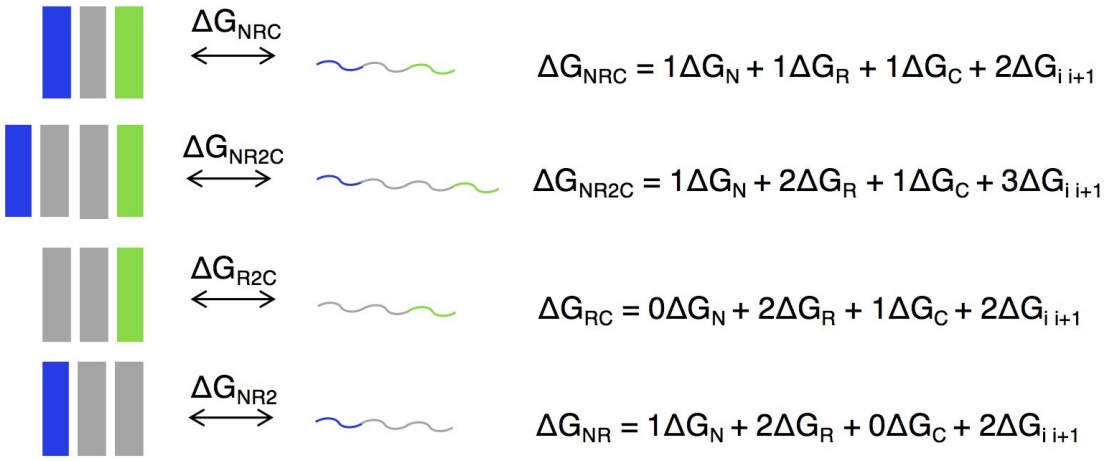


Figure 1.5 Ising analysis allows populations of rare microstates to be quantified. All possible microstates of a three repeat (N, R, and C) protein. In a one-dimensional Ising model, transitions between microstates are described by a linear combination of intrinsic and nearest-neighbor interfacial energies. Intrinsic energies describe the folding of each subunit (ΔG_N, ΔG_R, and ΔG_C) and interfacial energies describe formation of the interface between two adjacent folded repeats ($\Delta G_{i, i+1}$). Assuming the protein is linear and only adjacent subunits are directly coupled, the energetic transitions between all microstates are described above.

Figure 1.6 Length- and capping-dependence of stability. By studying the folding stability of consensus repeat arrays with variable capping identities and number of repeats, intrinsic and interfacial energies can be determined. In the first example, the free energy of complete unfolding of a protein with one N- cap repeat, one central R repeat, and one C-cap repeat includes one N-cap intrinsic energy, one central R intrinsic energy, one C-cap intrinsic energy, and two interfacial energies. Counting the number of a repeat type as well as the total number of interfaces gives similar free energy equations for other constructs. By including constructs with both caps, and either cap, as well as constructs with differing numbers of central repeats, a system is generated with enough unique equations to solve for the four unknown free energies (the intrinsic and interfacial energies). Although this analysis is useful to evaluate whether a series of proteins provides enough information to solve for free energy coefficients, a better numerical determination of these coefficients of obtained by fitting a statistical mechanical model that includes populations of partly folded states. More details of the statistical mechanics model are discussed in Chapter 2 and 3.



$$\Delta G_{NRC} = 1\Delta G_N + 1\Delta G_R + 1\Delta G_C + 2\Delta G_{i\,i+1}$$

$$\Delta G_{NR2C} = 1\Delta G_N + 2\Delta G_R + 1\Delta G_C + 3\Delta G_{i\,i+1}$$

$$\Delta G_{RC} = 0\Delta G_N + 2\Delta G_R + 1\Delta G_C + 2\Delta G_{i\,i+1}$$

$$\Delta G_{NR} = 1\Delta G_N + 2\Delta G_R + 0\Delta G_C + 2\Delta G_{i\,i+1}$$

$$\begin{bmatrix} \Delta G_{NRC} \\ \Delta G_{NR2C} \\ \Delta G_{R2C} \\ \Delta G_{NR2} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 2 \\ 1 & 2 & 1 & 3 \\ 0 & 2 & 1 & 2 \\ 1 & 2 & 0 & 2 \end{bmatrix} \begin{bmatrix} \Delta G_N \\ \Delta G_R \\ \Delta G_C \\ \Delta G_{i,i+1} \end{bmatrix}$$

CHAPTER 2

The unusual stability distributions of de novo designed helical repeat arrays: extreme global stability is determined by short-range interactions.

This chapter includes contributions for Kevin Sforza and Max Yuhas in the Barrick lab, and is a collaboration with David Baker and Fabio Parmegianni at University of Washington.

2.1 Abstract

Designed Helical Repeats (DHRs) are modular helix-loop-helix-loop protein structures that are tandemly repeated to form a superhelical array. Structures combining tandem DHRs demonstrate a wide range of molecular geometries, many of which are not observed nature. Understanding cooperativity of DHR proteins provides insight into the molecular origins of Rosetta-based protein design hyper-stability, and facilitates comparison of energy distributions in artificial and naturally occurring protein folds. Here we use a nearest-neighbor Ising model to quantify the intrinsic and interfacial free energies of four different DHRs. We find that unlike naturally occurring repeat proteins, the individual repeats of DHR proteins are intrinsically stable.

This high intrinsic stability for designed helical repeats has a number of important implications. First, unlike naturally occurring repeat proteins, favorable intrinsic folding adds to stabilizing interfaces, resulting in extraordinary thermostability. Second, the favorable intrinsic stability of DHRs should result in low kinetic barriers to folding and a downhill energy landscape for folding; thus,

DHR proteins should have very fast but potentially complex folding rates. Third, the intrinsic stability differences suggest that part of the success of Rosetta-based design results from capturing favorable local interactions. Finally, the intrinsic stability of DHRs may provide the energetic flexibility to mix different DHR types on one polypeptide chain, significantly expanding the repertoire of folded DHRs for applications such involving molecular recognition.

2.2 Introduction

Linear repeat proteins have proven to be useful model systems in the quest to better understand protein folding thermodynamics. Due to their repetitive primary structures, these proteins fold into linearly extended modular arrays with approximate translational symmetry. Unlike globular proteins, where interactions can span across the protein sequence, the interactions of linear repeat proteins are confined to within or between adjacent repeats[1]. This architecture simplifies the models used to describe protein folding thermodynamics, permitting the use of nearest-neighbor Ising analysis.

One dimensional Ising analysis has been successfully applied to a number of linear helical repeat proteins[2–5]. This analysis assumes that repeat protein stability can be parsed into intrinsic folding energies of individual repeats and coupling energies at the interfaces between adjacent folded repeats. Previous work characterizing linear repeat proteins derived from naturally-occurring folds shows that individual repeats are unstable. In these proteins, stability (and cooperativity) originates in the favorable interfaces between adjacent repeats.

The use of linear repeat proteins as molecular scaffold and recognition partners has been exploited in a number of engineering applications. Consensus ankyrin repeats have been used to select high affinity binding partners (Pluckthun refs) and to enhance the activity of engineered enzymes[6], transcription activator-like effector proteins (TALEs) have been engineered for in genome editing[7,8], and molecular chaperones have been fused to tetratricopeptide repeat proteins (TPRs) to increase substrate affinity[9]. While

repeat proteins designed from naturally occurring families have remarkable utility, expanding architectures beyond these folds would further enable such protein engineering applications. One promising set of templates are the *de novo* designed helical repeat proteins (DHRs)[10]. This series of constructs a wide variety of native-state architectures that extend beyond those of naturally occurring repeat proteins.

Here we characterize the stability of a series of DHRs using nearest-neighbor Ising analysis. We find that unlike naturally occurring repeat proteins, both the intrinsic and interfacial contributions to folding free energy of DHRs are thermodynamically favorable, giving rise to extraordinarily high folding stability while maintaining cooperativity. The favorable local stability of DHR repeats suggests a reduced folding barrier. The observation of favorable local stabilities in DHRs provides insights into the success of current Rosetta-based design, and suggests mechanisms for further DHR-based protein designs.

2.3 Results

Equilibrium unfolding of Designed Helical Repeat proteins

To investigate the thermodynamic folding behavior of Rosetta-designed repeat proteins with novel fold geometries, we chose DHR candidates for characterization based on the following criteria: (1) available SAXS and crystal structure data that demonstrate that the target structure is adopted, (2) an absence of cysteine residues to reduce complications associated with disulfide linkages, and (3) experimental evidence that shows the capped repeat proteins

to be monomeric in solution. The proteins DHR9, DHR10, DHR54, DHR71, and DHR79 (Figure 2.1A) satisfy these criteria.

Far-UV CD spectra for four repeat NR_2C constructs (where N and C represent N- and C-terminal polar capping repeats flanking two internal DHR repeats) for each of these DHRs display characteristic minima at 208 nm and 222 nm, consistent with folded α-helical proteins (Figure 2.1B).

To measure DHR stability, we monitored guanidine-HCl induced unfolding transitions using CD spectroscopy at 222nm. For DHR10, DHR54, DHR71, and DHR79, NR_2C constructs displayed a single sigmoidal unfolding transition, which is well-fitted with a two-state model for unfolding (Figure 2.1C). DHR9 did not unfold across a range of temperatures, pH, and denaturant concentrations (data not shown), precluding thermodynamic analysis. The unfolding transitions of DHRs 54, 71, and 79 have high slopes and midpoints for unfolding. The steep guanidine unfolding transitions of these three constructs suggest a high level of cooperativity. In contrast, the unfolding transition of DHR10.2 occurs over a broader range of denaturant concentration with a low midpoint compared to the other DHRs.

Length and capping dependence on stability

To determine the effects of variation in repeat number and the sequence substitutions associated with the N- and C-terminal capping repeats on stability, we constructed a series of DHR proteins that delete terminal and internal repeats For many singly-capped constructs, soluble oligomers could be detected by

sedimentation velocity analytical ultracentrifugation (SV-AUC). To eliminate oligomerization, glycerol was added to ten percent. SV-AUC demonstrates that in the presence of glycerol, most singly-capped constructs remain folded monomers (Figure 2.S1). For DHR10, deletion of the C-terminal repeat leads to formation of soluble oligomers even in the presence of glycerol. To prevent this oligomerization, we made a series of charged substitutions to solvent-exposed hydrophobic residues in the N-terminal capping repeat (V12K, I14E, V16E, L39R), we refer to this series as DHR10.2. All variants of DHR10.2 are monomeric.

For each of the four DHR series, we measured unfolding curves for constructs with two, three, and four repeats under conditions where constructs remain monomeric. Two repeat constructs contain a single R repeat with either an N-terminal capping repeat (NR) or a C-terminal capping repeat (RC). Three repeat constructs contain one construct with a single R repeat with both N- and C-terminal capping repeats (NRC), or two R repeats with either an N- (NR_2) or C-terminal (R_2C) capping repeat. The four repeat construct contains two R repeats with both N- and C-terminal capping repeats (NR_2C). For DHR54 we were also able to construct and characterize a folded, stable single N repeat.

Stabilities of length and capping variants were monitored by guanidine-HCl induced unfolding transitions by CD spectroscopy at 222nm as described above (Figure 2.2). For all DHR proteins, stability increases as the number of repeats increases (compare DHR54 N to NR and NR_2, DHRs 10.2, 71, 79 NR to

NR_2, and all DHRs NRC to NR_2C). However, the capping repeats are generally less stabilizing than internal "R" repeats.

For the DHR10.2 series, adding a C-terminal capping repeat to NR increases the transition slope and midpoint, whereas adding a C-terminal capping to NR_2 increases the slope more than midpoint (compare NR_2 to NR_2C). The C-terminal capping repeat gives rise to a larger slope and midpoint than the N-terminal capping repeat (compare NR_2 to R_2C), suggesting greater intrinsic stability for the C-cap, or a more stabilizing R:C interface.

For DHR54 and DHR71, the unfolding midpoint for N-terminal capped constructs are higher than those for C-terminal capped constructs (compare NR to RC). While for DHR54 capping identity does not affect transition slope, adding a C-terminal capping repeat to DHR71 appears to result in multistate unfolding behavior (compare NR to NRC, and NR_2 to NR_2C). Moreover, the N-cap repeat shifts the unfolding transition of DHR54 to higher guanidine concentration (compare NR to RC).

Ising analysis extracts intrinsic and interfacial folding free energies for all DHRs in the absence of glycerol

Intrinsic and interfacial folding energies were determined using a 1-D Ising model. In this model, individual repeats are monitored as either folded or unfolded states. Thus, for an *n*-repeat array, there are 2^n configurations treated by the model. The energy of each configuration is determined by the intrinsic

folding energy of each repeat (ΔG_i) as well as the coupling ("interfacial") free energies ($\Delta G_{i,i+1}$) between consecutive repeats.

Because the sequences of the N- and C-terminal capping repeats differ from the sequence of central repeats, three intrinsic energies are included in the model (ΔG_N, ΔG_R, and ΔG_C). For all DHRs except DHR54, the model includes only one interfacial free energy ($\Delta G_{i,i+1}$). Although it is possible that the free energies between central repeats and capping repeats differ, it is not possible to resolve such differences unless the unfolding energy of the lone cap can be measured. Because an unfolding transition of a lone N-cap repeat for DHR54 is observed, a separate term for the interfacial energy between an N-cap repeat and the adjacent central repeat ($\Delta G_{N,i+1}$) can be fitted.

To account for effects of glycerol on stability, we expanded our standard single-denaturant model to include a linear free intrinsic energy dependence on glycerol. This model was fitted to DHR guanidine-induced unfolding transitions collected at several glycerol concentrations[2,3,11]. By including guanidine HCl unfolding transitions at different glycerol concentrations, we were able to extract the intrinsic (ΔG_i) and interfacial ($\Delta G_{i,i+1}$) free energies in the absence of glycerol. For DHR10.2, DHR54, and DHR79, we assumed that N-cap, central, and C-cap repeats have identical m-values. For DHR71, fitting required a separate $m_{Gdn\text{-}HCl}$ for the C-cap repeat.

Figure 2.2 shows four global fits of the Ising model to DHR unfolding transitions. There are only six global thermodynamic parameters for the fits in Figure 2.2A and 2.2D and seven global thermodynamic parameters in Figures

2C and 2D. Global fits also include separate baseline parameters for each unfolding transition. For all DHR series, the data are well-fitted by the Ising model, and result in low and fairly random residuals. The largest non-random residuals are associated with the rather long native baselines associated with some of the longer constructs.

All DHRs have favorable interfacial free energies, similar to interfacial energies seen for naturally occurring repeat-proteins (refs). The intrinsic folding energies of DHRs are also favorable, in contrast with those of naturally occurring repeats. DHR71 and DHR10.2 have capping intrinsic energies that are unfavorable, consistent with the multi-state transitions seen in panel 2A and 2C. For all DHRs, glycerol is stabilizing, although the effects of glycerol on stability are significantly lower (and somewhat variable among DHR series) than that of guanidine HCl on a molar basis.

2.4 Discussion

By measuring the length-, capping-, and glycerol-dependence on stability of four DHRs families, we have used a 1D-Ising model to quantify their intrinsic folding free energies and interfacial coupling free energies. Unlike previously-studied helical repeat proteins, which were based on naturally-occurring folds, these proteins were generated by *de novo* design. Quantifying the cooperativity of DHRs using the Ising approach provides a new vantage point to compare and contrast natural and designed proteins. The surprising finding that DHRs have intrinsically stable repeats has important implications for understanding the

energetic basis for the success in Rosetta design, for the distribution of cooperativity in naturally occurring repeat proteins, and for the kinetics of folding as a barrier-limited versus downhill process.

Rosetta algorithms design stable proteins through favorable local interactions

In the past decade, 1D Ising analysis has been used to dissect folding cooperativity in a variety of naturally-occurring helical repeat protein families[2–5,11]. These proteins have typically been designed using consensus information obtained from multiple sequence alignments, although for some of these series[4,5], designs were based on genes with nearly identical sequence repeats. Although exact numbers vary, all of these naturally occurring repeat proteins have favorable interfacial (i.e., negative) free energies between repeats (unfilled blue circles, Figure 2.3A), which are partly offset by unfavorable (positive) intrinsic folding free energies (unfilled red circles, Figure 2.3B).

The interfacial energies between the designed helical repeats are also stabilizing, and span roughly the same range as interfacial energies of the naturally-occurring repeat proteins. However, fitted DHR intrinsic folding energies are favorable (Figure 2.3A), in contrast to all previously measured intrinsic energies for natural repeat proteins [3,11,2,4,5]. This enhancement to stability of intrinsic free energies may reflect a fundamental difference between Rosetta-based *de novo* design[10] and natural selection. It appears that Rosetta-based design is particularly good at enhancing local stability. Whether this

enhancement results from backbone selection in the early stages of design, sequence design in the intermediate stages, or selection for funneled energy landscapes is unclear. In contrast, the *de novo* design protocol used for DHRs appears to be about the same as natural selection in stabilizing interfaces between repeats.

One consequence of the uniquely stabilizing intrinsic folding energies seen for DHRs is that they significantly enhance overall stability. The stability of a tandem repeat array depends on both the intrinsic and interfacial stabilities. The sum of the intrinsic and interfacial free energies gives the stability increment of adding a repeat to an existing folded array (Figure 2.3C). For naturally-occurring repeat proteins, this stability increment derives solely from the interfacial interaction energy, and is offset by the intrinsic energy. For DRH arrays, the favorable intrinsic folding energies add to the interfacial energies, giving rise to an exceptionally large free energy decrease for adding a repeat to an existing array. This results in very high native-state stabilities.

Differences between the energy landscapes of de novo designed and naturally-occurring helical repeat proteins. Quantification of the intrinsic and interfacial free energies of repeat proteins using the Ising model allows the energy landscapes of repeat proteins to be represented in meaningful reaction coordinates and with experimentally determined free energies[12,13]. In this representation, the free energies of states where one or more adjacent repeats are folded and paired are plotted as a function of the number of folded repeats

and the location of the partly folded structure (N-terminal, C-terminal, or internal; Figure 2.4A). Ignoring lower probability configurations where unfolded repeats are flanked by folded repeats, there are ten configurations depicted in the NR_2C landscape. Because the intrinsic folding energies of naturally-derived consensus ankyrin repeats are unfavorable, all conformations with one folded repeat have high energies, resulting in a large barrier that must be crossed during folding. Depending on the structure of the transition state for folding, even higher barriers in which a second ankyrin repeat is at least partly folded[14] but not yet paired may further impede folding. In contrast, because the intrinsic folding energy of DHR54 repeats designed repeats is favorable, all partly-folded configurations are lower in energy than the native state under conditions that strongly stabilize folding. Thus the energy landscape for DHR54 folding is comparatively smooth and downhill. Moreover, since addition of each folded DHR54 repeat significantly decreases the free energy, the landscape is also very steep, reflecting a strong driving force for folding. Given these landscape features, DHR54 should fold much faster than cAnk.

Unstable repeats may be a result of natural selection for folding cooperativity.

In addition to reflecting successful design principles for DHRs, the difference between intrinsic stabilities of natural and designed helical repeats

may reflect features imposed by natural selection on natural repeat folds. Instability of local repeats enhances cooperativity, suppressing both the equilibrium formation of partly folded states and the transient formation of partly structured species through a zippering mechanism during folding. Such concentrations may be prone to misfolding and aggregation. Naturally occurring repeat proteins may have evolved to minimize such structures by partitioning stability into interfacial interactions rather than intrinsic folding of repeats. Obviously, there is no such pressure on designed helical repeats. This proposal is consistent with ideals that have emerged from energy landscape theory that natural proteins have been selected to minimize energetic frustration[15–19]. Moreover, family-specific functional constraints on naturally-occurring repeat proteins may modulate cooperativity to allow for precise conformational fluctuations, as has been suggested for DNA-binding by TALE-repeat proteins[5].

Lastly, it is possible that nature doesn't select for or against unfavorable intrinsic energies in repeat proteins. Because repeat proteins have very favorable interfacial free energies, global stability is achieved in the presence or absence of stabilizing intrinsic energy. Maybe the selection pressure is to generate proteins above a threshold of global stability, and it is easier to maintain a few very stabilizing interfacial interactions[20,21] while allowing for functional sequence variation that decreases intrinsic energy.

2.5 Methods

Cloning, expression, and purification

Genes containing DHR repeat constructs were purchased as GeneStrings from GeneArt and cloned with C-terminal His_6 tags via Gibson Assembly. DHR constructs were grown in BL21(T1R) cells at 37°C to an OD of 0.6-0.8, induced with 0.2 mM IPTG, and expressed overnight at 17°C. Following cell pelleting, resuspension, and lysis, proteins were purified by affinity chromatography on an Ni-NTA column. Proteins were eluted using 250 mM imidazole and dialyzed into 150 mM NaCl, 0-20% glycerol, and 25 mM $NaPO_4$ pH 7.0.

Circular Dichroism (CD) spectroscopy

Circular Dichroism measurements were collected using an AVIV model 400 CD Spectrometer (Aviv Associates, Lakewood, NJ, USA). Far-UV CD scans were collected at 25°C using an 0.1 cm pathlength quartz cuvette, with protein concentrations of 15-30 μM. Buffer scans were recorded and were subtracted from the raw CD data. CD-monitored guanidine unfolding transitions at 222 nm were generated with an automated titrator using 1.5-3 μM protein and a 1 cm pathlength quartz cuvette.

Ising analysis

To determine the intrinsic and interfacial free energies for folding of DHR arrays, and to analyze energies of partly folded states, we used a one-dimensional Ising formalism [22,23]. In this model, intrinsic folding and interfacial interaction between nearest neighbors are represented using equilibrium constants κ and τ, respectively, where

$$\kappa_N = e^{-\left(\Delta G_N - m_{GdmHCl}[GdmHCl] - m_{glycerol}[glycerol]\right)/RT} \quad (1)$$

$$\kappa_R = e^{-\left(\Delta G_R - m_{GdmHCl}[GdmHCl] - m_{glycerol}[glycerol]\right)/RT} \quad (2)$$

$$\kappa_C = e^{-\left(\Delta G_C - m_{GdmHCl}[GdmHCl] - m_{glycerol}[glycerol]\right)/RT} \quad (3)$$

$$\tau = e^{-\left(\Delta G_{R,\,i-1}\right)/RT} \quad (4)$$

For all DHRs, the intrinsic folding free energies of *N* (solubilizing N-terminal cap), *R* (consensus repeat), and *C* (solubilizing C-terminal cap) are independent adjustable parameters. DHR10.2, DHR71, and DHR79 are well described by a simple model where the interfacial interactions of the *N:R* , *R:R*, and *R:C* pairs are identical. DHR54 unfolding transitions are better fitted by a model where the interfacial interactions of the *R:R* and *R:C* interface are identical, whereas the *N:R* pair is different. Glycerol and GdnHCl dependences are built into the intrinsic (but not the interfacial) terms. DHR71 unfolding transitions are better fitted by a model that includes a separate denaturant dependence for the C-terminal cap ($m_{GdnHCl,\,C}$, Table 1).

Using these equilibrium constants, a partition function *q* for an *n*-repeat construct can be constructed by multiplying two-by-two transfer matrices:

$$q = \begin{bmatrix} 0 & 1 \end{bmatrix} \begin{bmatrix} \kappa_N\tau & 1 \\ \kappa_N & 1 \end{bmatrix} \begin{bmatrix} \kappa_R\tau & 1 \\ \kappa_R & 1 \end{bmatrix}^{n-2} \begin{bmatrix} \kappa_C\tau & 1 \\ \kappa_C & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad (2)$$

This representation correlates the each repeat to its neighbor through the separate rows of each matrix. The fraction of folded protein (f_{folded}) can be obtained by differentiation:

$$f_{folded} = \frac{1}{nq}\left(\kappa_N \frac{\partial q}{\partial \kappa_N} + \kappa_R \frac{\partial q}{\partial \kappa_R} + \kappa_C \frac{\partial q}{\partial \kappa_C}\right) \quad (3)$$

Ising parameters were determined by globally fitting Eq. 3 to guanidine-induced unfolding transitions collected at 0, 10, and 20% glycerol. Fitting was performed using the nonlinear least squares algorithm of the lmfit package[24] using an in-house python program (written by J. Marold[4] and adapted to include glycerol dependence by K.G.-S.) Confidence intervals (95%) were determined by performing 2000 bootstrap iterations.

Acknowledgements

The authors thank members of the Barrick and Baker lab for their input on this work, Jeff Gray for helping to initiate these studies. The authors acknowledge the support of the Center for Molecular Biophysics at Johns Hopkins and Dr. Katherine Tripp for instrumental and technical support.

2.6 References

1. Kloss, E., Courtemanche, N. & Barrick, D. Repeat-protein folding: new insights into origins of cooperativity, stability, and topology. *Arch. Biochem. Biophys.* **469,** 83–99 (2008).
2. Kajander, T., Cortajarena, A. L., Main, E. R. G., Mochrie, S. G. J. & Regan, L. A new folding paradigm for repeat proteins. *J. Am. Chem. Soc.* **127,** 10188–10190 (2005).
3. Aksel, T., Majumdar, A. & Barrick, D. The contribution of entropy, enthalpy, and hydrophobic desolvation to cooperativity in repeat-protein folding. *Struct. Lond. Engl. 1993* **19,** 349–360 (2011).
4. Marold, J. D., Kavran, J. M., Bowman, G. D. & Barrick, D. A Naturally Occurring Repeat Protein with High Internal Sequence Identity Defines a New Class of TPR-like Proteins. *Struct. Lond. Engl. 1993* (2015). doi:10.1016/j.str.2015.07.022
5. Geiger-Schuller, K. & Barrick, D. Broken TALEs: Transcription Activator-like Effectors Populate Partly Folded States. *Biophys. J.* **111,** 2395–2403 (2016).
6. Cunha, E. S., Hatem, C. L. & Barrick, D. Synergistic enhancement of cellulase pairs linked by consensus ankyrin repeats: Determination of the roles of spacing, orientation, and enzyme identity. *Proteins* **84,** 1043–1054 (2016).
7. Christian, M. *et al.* Targeting DNA Double-Strand Breaks with TAL Effector Nucleases. *Genetics* **186,** 757–761 (2010).
8. Li, T. *et al.* TAL nucleases (TALNs): hybrid proteins composed of TAL effectors and FokI DNA-cleavage domain. *Nucleic Acids Res.* **39,** 359–372 (2011).

9. Cortajarena, A. L., Kajander, T., Pan, W., Cocco, M. J. & Regan, L. Protein design to understand peptide ligand recognition by tetratricopeptide repeat proteins. *Protein Eng. Des. Sel. PEDS* **17,** 399–409 (2004).
10. Brunette, T. J. *et al.* Exploring the repeat protein universe through computational protein design. *Nature* **528,** 580–584 (2015).
11. Wetzel, S. K., Settanni, G., Kenig, M., Binz, H. K. & Plückthun, A. Folding and unfolding mechanism of highly stable full-consensus ankyrin repeat proteins. *J. Mol. Biol.* **376,** 241–257 (2008).
12. Mello, C. C. & Barrick, D. An experimentally determined protein folding energy landscape. *Proc. Natl. Acad. Sci. U. S. A.* **101,** 14102–14107 (2004).
13. Aksel, T. & Barrick, D. Direct observation of parallel folding pathways revealed using a symmetric repeat protein system. *Biophys. J.* **107,** 220–232 (2014).
14. Ferreiro, D. U., Cho, S. S., Komives, E. A. & Wolynes, P. G. The energy landscape of modular repeat proteins: topology determines folding mechanism in the ankyrin family. *J. Mol. Biol.* **354,** 679–692 (2005).
15. Bryngelson, J. D. & Wolynes, P. G. Spin glasses and the statistical mechanics of protein folding. *Proc. Natl. Acad. Sci. U. S. A.* **84,** 7524–7528 (1987).
16. Bryngelson, J. D., Onuchic, J. N., Socci, N. D. & Wolynes, P. G. Funnels, pathways, and the energy landscape of protein folding: a synthesis. *Proteins* **21,** 167–195 (1995).

17. Onuchic, J. N., Luthey-Schulten, Z. & Wolynes, P. G. Theory of protein folding: the energy landscape perspective. *Annu. Rev. Phys. Chem.* **48,** 545–600 (1997).
18. Onuchic, J. N. & Wolynes, P. G. Theory of protein folding. *Curr. Opin. Struct. Biol.* **14,** 70–75 (2004).
19. Oliveberg, M. & Wolynes, P. G. The experimental survey of protein-folding energy landscapes. *Q. Rev. Biophys.* **38,** 245–288 (2005).
20. Preimesberger, M. R. *et al.* Direct NMR detection of bifurcated hydrogen bonding in the α-helix N-caps of ankyrin repeat proteins. *J. Am. Chem. Soc.* **137,** 1008–1011 (2015).
21. Kloss, E. & Barrick, D. C-terminal deletion of leucine-rich repeats from YopM reveals a heterogeneous distribution of stability in a cooperatively folded protein. *Protein Sci. Publ. Protein Soc.* **18,** 1948–1960 (2009).
22. Aksel, T. & Barrick, D. Analysis of repeat-protein folding using nearest-neighbor statistical mechanical models. *Methods Enzymol.* **455,** 95–125 (2009).
23. Poland, D. & Scheraga, H. A. *Theory of helix-coil transitions in biopolymers: statistical mechanical theory of order-disorder transitions in biological macromolecules*. (Academic Press, 1970).
24. Newville, M., Stensitzki, T., Allen, D. B. & Ingargiola, A. *LMFIT: Non-Linear Least-Square Minimization and Curve-Fitting for Python*¶. (Zenodo, 2014).

25. Schuck, P. Size-distribution analysis of macromolecules by sedimentation velocity ultracentrifugation and lamm equation modeling. *Biophys. J.* **78,** 1606–1619 (2000).
26. John Philo's Software Home Page. Available at: http://www.jphilo.mailway.com/. (Accessed: 6th October 2017)

Table 2.1. Thermodynamic parameters obtained from Ising fits.

	ΔG_N	ΔG_R	ΔG_C	$\Delta G_{i,i+1}$	$m_{GdnHCl,\ i}$	$m_{Glycerol,\ i}$	$m_{GdnHCl,\ C}$	$\Delta G_{N,i+1}$
DHR10.2	1.46 [1.26, 1.67]	-2.51 [-2.90, -2.15}	0.63 [0.32, 1.00]	-4.80 [-5.10, -4.53]	-1.23 [-1.33, -1.14]	0.36 [0.33, 0.40]	N/A	N/A
DHR54	-0.45 [-0.58, -0.32]	-2.04 [-2.17, -1.92]	-0.84 [-0.94, -0.74]	-6.76 [-6.98, -6.54]	-1.24 [-1.28, -1.21]	0.41 [0.39, 0.43]	N/A	-7.72 [-7.95, -7.49]
DHR71	-3.01 [-3.27, -2.75]	-1.41 [-1.61, -1.23]	3.06 [2.87, 3.29]	-9.93 [-10.50, -9.43]	-1.57 [-1.66, -1.49]	0.17 [0.15, 0.20]	-0.71 [-0.79, -0.64]	N/A
DHR79	-1.84 [-2.06, -1.64]	-3.48 [-3.83, -3.22]	-1.81 [-2.08, -1.61]	-4.83 [-5.14, -4.55]	-1.12 [-1.18, -1.06]	0.15 [0.12, 0.18]	N/A	N/A

Free energies have units of kcal/mol. m_{GdnHCl} and $m_{Glycerol}$ have units of kcal/mol/[M GdnHCl] and kcal/mol/[M Glycerol]. 95% confidence intervals shown in brackets are from 2,000 iterations of bootstrap analysis.

Figure 2.1. Structures and stabilities of designed helical repeat proteins. (A) Selected DHR proteins have distinct structures that are to-date unobserved in natural repeat proteins, including unique inter-repeat twists and radii of curvature between repeating units. (B) Far-UV circular dichroism shows characteristic α-helical spectra for DHR proteins. (C) Guanidine-induced unfolding transitions of four-repeat NR_2C DHR proteins (red circles) fit with a two-state unfolding model (black curves) reveal stable, cooperative folding behavior of the DHR proteins. Panels in (B) and (C) correspond to the DHR proteins shown in (A). PDB codes are 5CWG (DHR10), 5CWL), DHR71 (DHR54), 5CWN (DHR71), and 5CWP (DHR79).

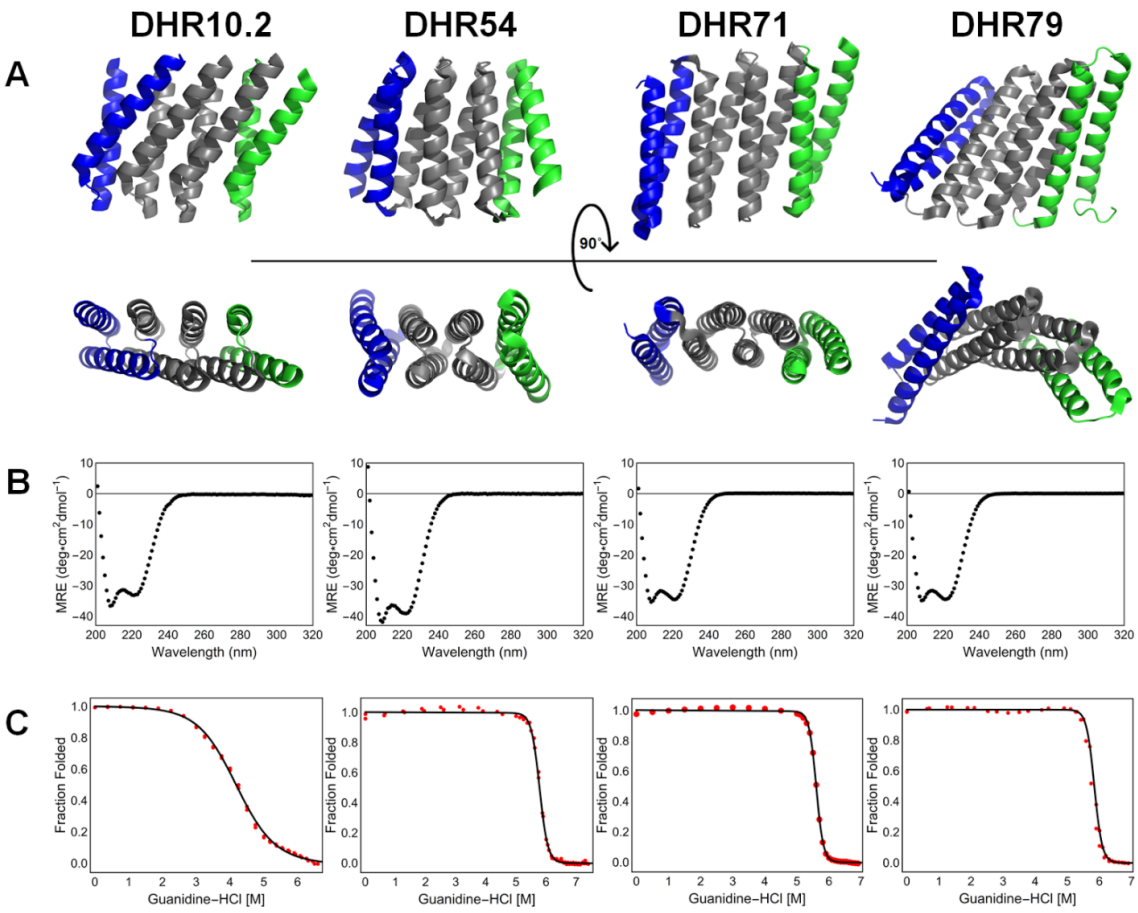

A
DHR10.2
DHR54
DHR71
DHR79
90°
B
MRE (deg•cm²dmol⁻¹)
Wavelength (nm)
C
Fraction Folded
Guanidine–HCl [M]

Figure 2.2. Unfolding transitions and nearest-neighbor Ising analysis of DHR proteins of different length and capping architecture. Guanidine-induced unfolding transitions were fitted with a nearest-neighbor Ising model (curves). N-capped constructs are shown in blue, C-capped constructs are shown in grey, and doubly-capped constructs are shown in red. Differences in glycerol concentrations are shown using different line styles: 0% glycerol, dash-dotted curves; 10% glycerol, solid curves; 20% glycerol, dashed curves). For all constructs, increasing the number of repeats increases stability (based un unfolding midpoints). Likewise, increasing glycerol concentration increases stability, although glycerol stabilizes DHR10.2 (A) and DHR54 (B) to a greater extent than DHR71 (C) and DHR79 (D). Conditions: 25 mM $NaPO_4$, 150 mM NaCl, 25°C.

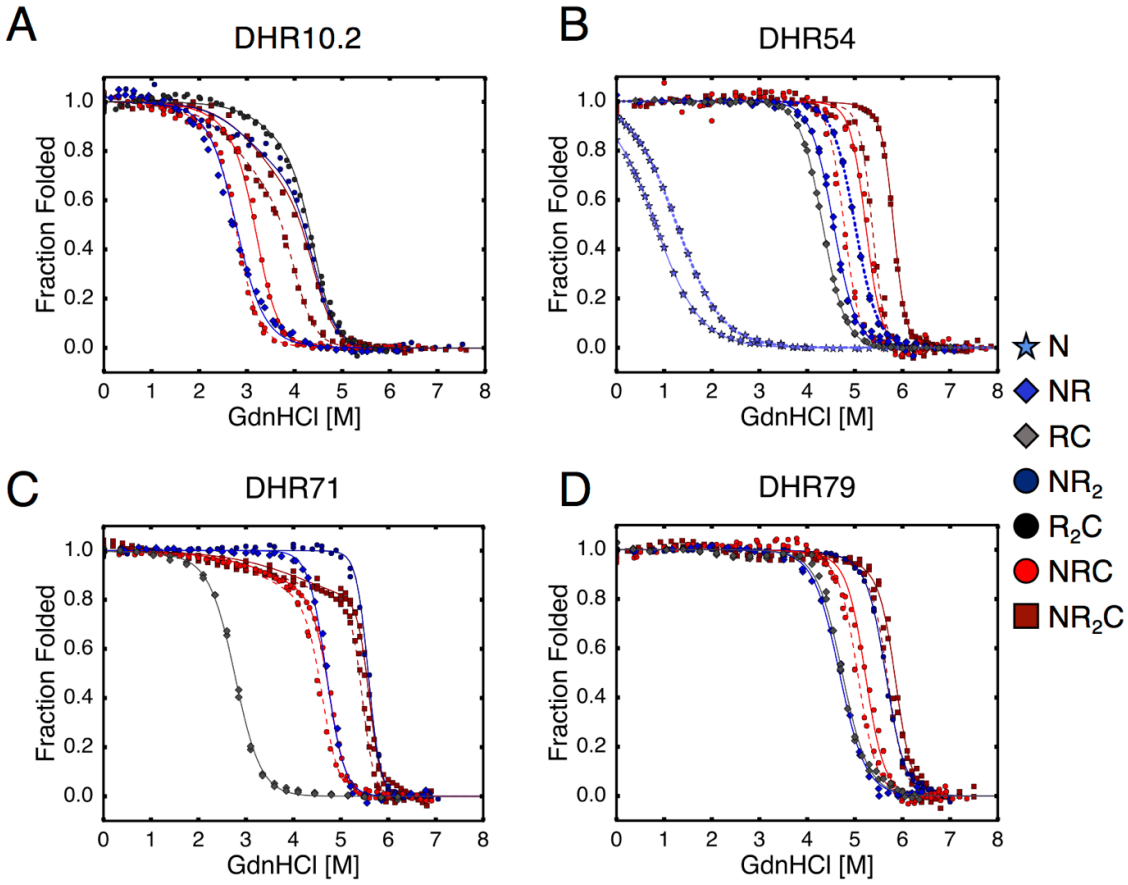

A
DHR10.2
B
DHR54
C
DHR71
D
DHR79
Fraction Folded
GdnHCl [M]
N
NR
RC
NR2
R2C
NRC
NR2C

Figure 2.3. DHR repeats are intrinsically stable, unlike the repeats of naturally occurring repeat proteins. (A) Intrinsic and (B) interfacial coupling free energies determined by Ising analysis for designed helical repeat proteins (filled circles, this study) and natural repeat proteins (open circles, $TALES^{NS}$ and $TALES^{HD}$ [5], 42PR [4], cANK [3], cTPR [4]). Unfavorable (i.e., positive) free energy terms are in red, favorable (i.e., negative) folding free energies are in blue. Designed helical repeats are stabilized by both favorable intrinsic and interfacial coupling folding free energies, while natural repeat proteins are destabilized by unfavorable intrinsic folding free energies, which partly offset favorable interfacial interactions. (C) Free energy associated with adding a single repeat to a folded array (the sum of free intrinsic and interfacial free energies in panels A and B). Due to both their favorable intrinsic folding free energies, DHR proteins are more strongly stabilized by the addition of repeats than natural repeat proteins, and as a result, are extraordinarily stable.

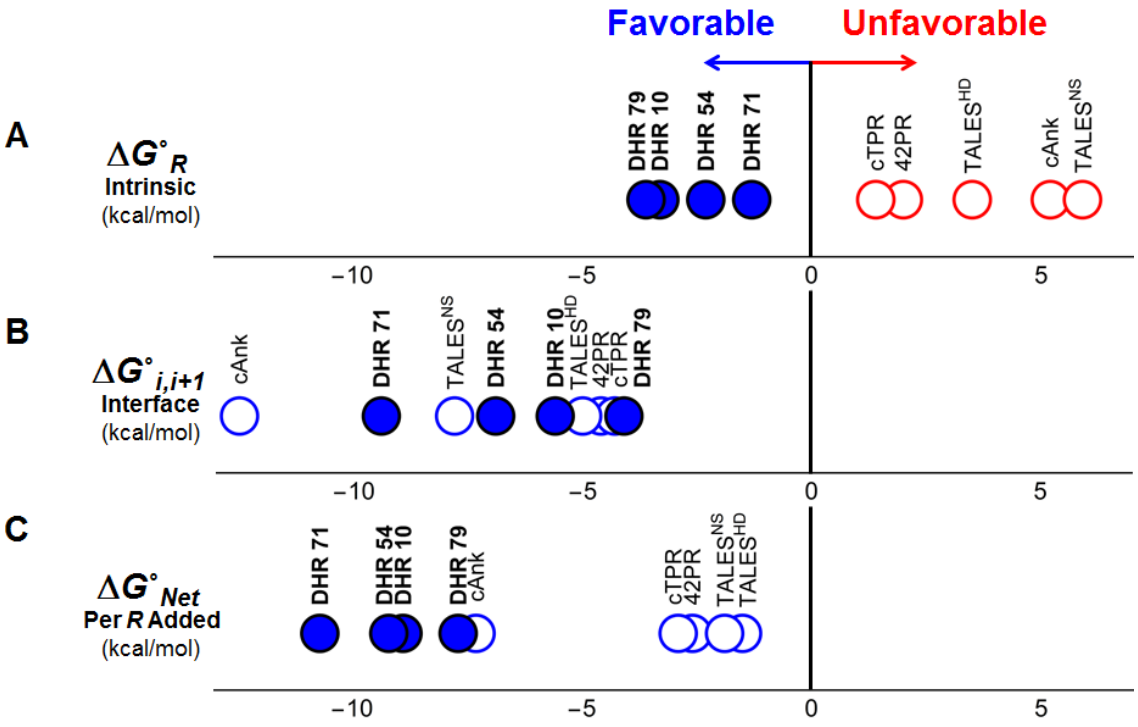
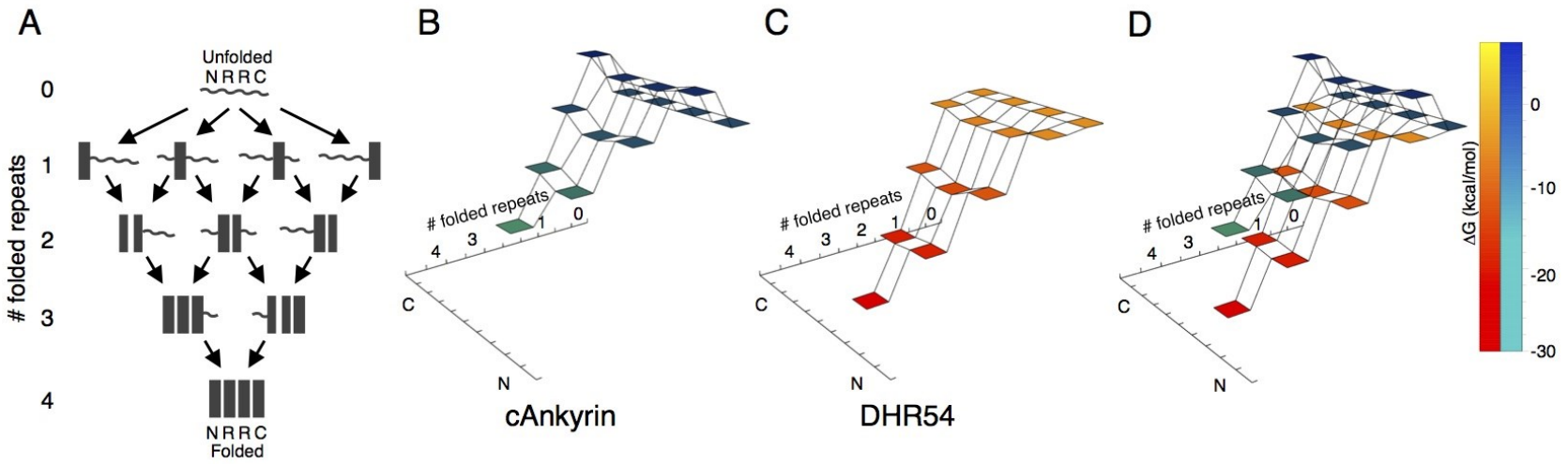

Favorable
Unfavorable
A
ΔG°R
Intrinsic
(kcal/mol)
DHR 79
DHR 10
DHR 54
DHR 71
cTPR
42PR
TALES^HD
cAnk
TALES^NS
−10
−5
0
5
B
ΔG°i,i+1
Interface
(kcal/mol)
cAnk
DHR 71
TALES^NS
DHR 54
DHR 10
TALES^HD
42PR
cTPR
DHR 79
−10
−5
0
5
C
ΔG°Net
Per R Added
(kcal/mol)
DHR 71
DHR 54
DHR 10
DHR 79
cAnk
cTPR
42PR
TALES^NS
TALES^HD
−10
−5
0
5

Figure 2.4. Stabilizing intrinsic energies create barrierless folding energy landscapes for DHR proteins in the absence of denaturant. (A) Repeat proteins with NR_2C repeat sequences can fold along many pathways. (B and C) Free energy landscapes from experimentally determined intrinsic and interfacial free energies. The vertical dimension (and shading) shows the free energies of partly folded states along the folding pathway shown in (A). (B) Consensus ankyrin repeat proteins, which are based on the naturally occurring ankyrin repeat family, have destabilizing intrinsic energies, and as a result, folding the first repeat results in a barrier to folding. (C) DHR54 proteins have stabilizing intrinsic folding energies, folding the first repeat is energetically favorable, and addition of subsequent repeats are strongly downhill. Landscapes were generated with Mathematica, using the "Polygon" primitive of the "Graphics3D" command. (D) An overlay of cAnkyrin (blue-green) and DHR54 (orange-red) free energy landscapes highlighting unique features of each landscape.


A
Unfolded
N R R C
0
1
2
3
4
folded repeats
N R R C
Folded
B
folded repeats
4
3
1
0
C
N
cAnkyrin
C
folded repeats
4
3
2
1
0
C
N
DHR54
D
folded repeats
4
3
1
0
C
N
ΔG (kcal/mol)
0
-10
-20
-30

2.7 Supplemental Material

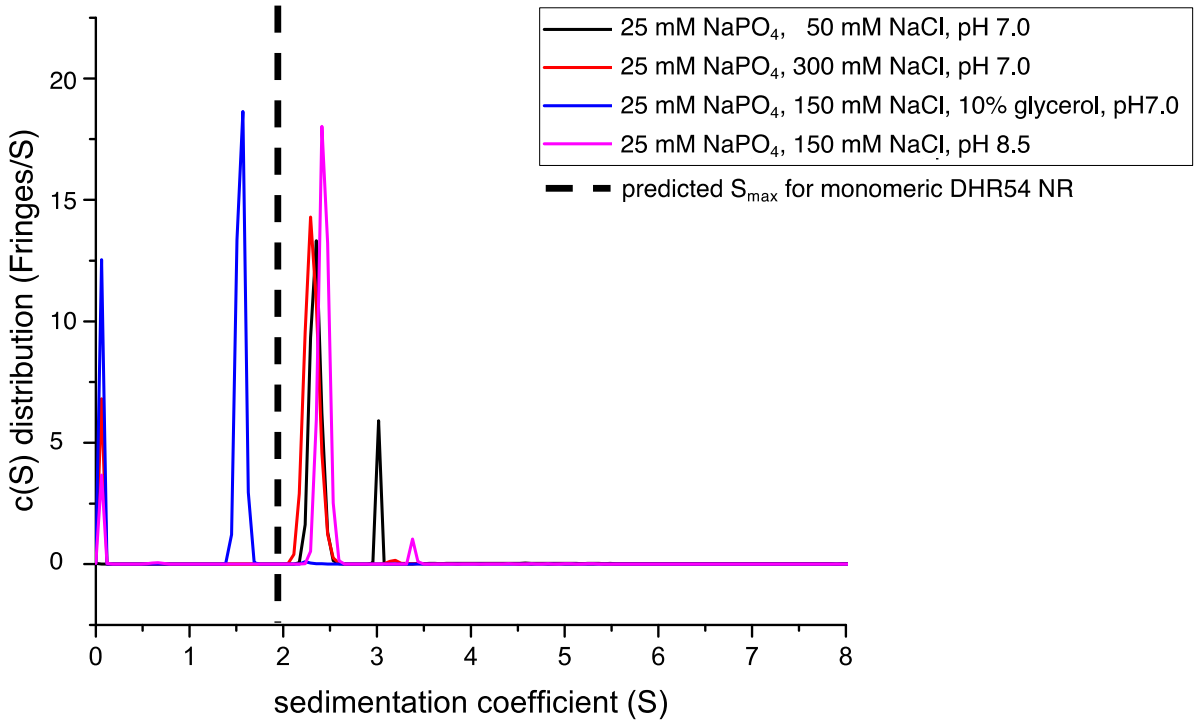


Figure 2.S1. Sedimentation Velocity c(S) plot for DHR54 NR in the absence and presence of glycerol. Sedimentation velocity experiments were performed with DHRs. Data were processed and fitted in Sedfit[25] as previously described[4]. The predicted Smax was calculated for DHR54 NR using Sednterp[26]. In the presence of 10 % glycerol, the c(S) distributions are consistent with monomers.

CHAPTER 3

Broken TALEs: Transcription Activator-Like Effectors (TALEs) populate partly folded states

The work in this chapter is published in the Biophysical Journal.

Authors: Kathryn Geiger-Schuller and Doug Barrick

3.1 Abstract

Transcription activator-like effector proteins (TALEs) contain large numbers of repeats that bind double stranded DNA, wrapping around DNA to form a continuous superhelix. Since unbound TALEs retain superhelical structure, it seems likely that DNA binding requires a significant structural distortion or partial unfolding. Here we use nearest-neighbor “Ising” analysis of consensus TALE (cTALE) repeat unfolding to quantify intrinsic folding free energies, coupling energies between repeats, and the free energy distribution of partly unfolded states, and to determine how those energies depend on the sequence that determines DNA-specificity (called the "RVD"). We find a moderate level of cooperativity for both the HD and NS RVD sequences (stabilizing interfaces combined with unstable repeats), as has been seen in other linear repeat proteins. Surprisingly, RVD sequence identity influences both the overall stability and the balance of intrinsic repeat stability and interfacial coupling energy.

Using parameters from the Ising analysis, we have analyzed the distribution of partly folded states as a function of cTALE length and RVD sequence. We find partly unfolded states where one or more repeats are

unfolded to be energetically accessible. Mixing repeats with different RVD sequences increases the population of partially folded states. Local folding free energies plateau for central repeats, suggesting that TALEs access partially folded states where a single internal repeat is unfolded while adjacent repeats remain folded. This breakage should allow TALEs to access superhelically-broken states, and may facilitate DNA binding.

3.2 Introduction

Transcription Activator-Like Effectors (TALEs) are bacterial proteins containing a domain of tandem 34-residue repeats that binds to specific DNA sequences and affect transcription of host genes (1, 2). TALE repeat domains have an average of 17.5 repeats, although this number varies (3). Repeats have high sequence identity, with most of the variability at repeat positions 12 and 13. These two residues, together called repeat variable diresidues (RVDs), impart DNA binding specificity, and identities at positions 12 and 13 can be used to design TALE proteins that bind specific DNA sequences (4–6). Using this specificity code, TALE nucleases (TALENs) have been engineered for genome editing purposes (7, 8). TALE repeats have also been used to design proteins that bind to DNA and activate or repress transcription (9–13), modify DNA by demethylation (14), and probe chromatin dynamics by encoding fluorophores with high sequence specificity (15, 16). One challenge to TALE-based genome editing is cloning difficulties resulting from the large number of repeats needed for affinity/specificity (17). Enhancing the affinity of TALE repeats would thus be beneficial for genome editing and molecular genetic studies, allowing high-specificity target recognition with fewer repeats.

TALEs in their unbound state are superhelical, with 11 repeats per turn (18), and are not likely to thread onto DNA easily. It seems likely that a conformational change is required for binding. One way to bring about such a change is a global deformation of the native state. Some degree of native-state plasticity is suggested from comparison of the free and bound states (Fig. 3.S1 in

Supporting Material), but does not seem to be enough to adequately open the structure. Alternatively, a binding competent state could be reached by a localized structural transition, either by disrupting one or more interfaces between repeats, and or by unfolding one or more repeats.

Ising analysis provides a direct means to determine the cooperativity of linear arrays of identical sequence repeats, and for quantifying the populations of partly folded states (19–22). By studying the length dependence of stability, 1-D Ising analysis can resolve the energetics of folding individual repeats from the energetics of formation of interfaces between repeats. Two types of repeat proteins have been subjected to Ising analysis (TPR (20, 23) and ankyrin repeats (19, 21, 22)). By applying 1-D Ising analysis to identical TALE repeat arrays of different lengths, we can resolve folding free energies into intrinsic and interfacial components, quantify the extent of cooperativity in folding, and determine the populations of partially folded states that may facilitate DNA binding.

Here, we characterize the equilibrium stability of a series of TALE constructs with varying length and RVD sequence using nearest-neighbor Ising analysis. The length dependence and Ising analysis demonstrate an intermediate level of coupling between repeats. Local folding free energies (ΔG°_{local}), calculated from intrinsic and interfacial free energies, suggest significant populations of partially unfolded states. Surprisingly, the extent of coupling and local folding free energy profiles depend on the sequence of the RVDs. This dependence leads to a stability switch between NS- and HD-containing TALE arrays at a length of 8 repeats. The stabilities of mixed NS and HD constructs

demonstrate that RVD sequence identity partitions asymmetrically into its N- and C-terminal interfaces, introducing further variation in local folding free energies.

3.3 Materials and Methods

Cloning, expression, and purification

Consensus TALE repeat constructs were cloned with C-terminal His_6 tags via an in-house version of Golden Gate cloning (24). TALE constructs were grown in BL21(T1R) cells at 37°C to an OD of 0.6-0.8 and induced with 1 mM IPTG. Following cell pelleting, resuspension, and lysis, proteins were purified by resuspending the insoluble material in 6M urea, 300 mM NaCl, and 10 mM Tris pH 7.4. Constructs were loaded onto an Ni-NTA column. Protein was eluted using 250 mM imidazole and refolded during dialysis into 300 mM NaCl, 5% glycerol, and 10 mM Tris pH 7.4.

Circular Dichroism (CD) spectroscopy

Circular Dichroism measurements were collected using an AVIV model 400 CD Spectrometer (Aviv Associates, Lakewood, NJ, USA). Far-UV CD scans were collected at 25°C using an 0.1 cm pathlength quartz cuvette, with protein concentrations of 15-30 μM. Buffer scans were recorded and were subtracted from the raw CD data.

Urea-induced unfolding transitions

For short constructs, CD-monitored unfolding titrations at 222 nm were generated with an automated titrator. For N-capped constructs six repeats or longer, and all constructs seven repeats or longer, slow relaxation kinetics prevented us from collecting automated titrations; thus, we performed manual urea titrations for

these constructs. Solutions containing 0 and 8 M urea, each with 2 μM protein, were combined in various proportions using Hamilton syringes. Samples equilibrated for 12-24 hours at room temperature before monitoring CD signal at 222 nm.

Ising analysis

To determine the intrinsic and interfacial free energies for folding of cTALE arrays, and to analyze energies and populations of partly folded states, we used a one-dimensional Ising formalism (25, 26). In this model, intrinsic folding and interfacial interaction between nearest neighbors are represented using equilibrium constants κ and τ, respectively, where

$$\kappa_N = e^{-(\Delta G_N - 4m[urea])/RT} \quad (1.1)$$

$$\kappa_R = e^{-(\Delta G_R - m[urea])/RT} \quad (1.2)$$

$$\kappa_R = e^{-(\Delta G_C - m[urea])/RT} \quad (1.3)$$

$$\tau_N = e^{-(\Delta G_{N,\,i+1})/RT} \quad (1.4)$$

$$\tau_R = e^{-(\Delta G_{R,\,i+1})/RT} \quad (1.5)$$

In the current analysis, the intrinsic folding free energies of *N* (solubilizing N-terminal cap), *R* (consensus repeat), and *C* (solubilizing C-terminal cap) are each considered to be unique. The interfacial interactions of the *R:R* and *R:C* pairs are considered to be identical, but that the *N:R* pair is considered to be unique. Denaturant dependences are built into the intrinsic (but not the interfacial) terms.

The urea dependence of the *N*-terminal cap is scaled by a factor of 4 because there are four TALE-like repeats in the N-terminal cap.

Using these equilibrium constants, a partition function *q* can be constructed for an *n*-repeat construct by multiplying two-by-two transfer matrices as follows:

$$q = \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} \kappa_N \tau_N & \kappa_N \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \kappa_R \tau_R & \kappa_R \\ 1 & 1 \end{bmatrix}^{n-2} \begin{bmatrix} \kappa_C \tau_R & \kappa_C \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} \quad (2)$$

This representation differs from several previous expressions in that the matrices in Eq. 2 correlate to the next repeat (rather than the previous repeat)(25). Though this rephrasing does not alter *q*, it associates the κ and τ terms for the same repeat. The fraction folded (f_{folded}) can be determined.

$$f_{folded} = \frac{1}{nq}\left(\kappa_N \frac{\partial q}{\partial \kappa_N} + \kappa_R \frac{\partial q}{\partial \kappa_R} + \kappa_C \frac{\partial q}{\partial \kappa_C} \right) \quad (3)$$

A specific example for a three-repeat cTALE construct is given in the Supporting Material (Fig. 3.S2).

Ising parameters were determined by nonlinear least squares using an in-house python program (written by J. Marold)(23) by globally fitting Eq. 3 to urea-induced unfolding transitions. Confidence intervals were determined by performing 2000 bootstrap iterations (95%).

Calculation of local folding free energies within TALE arrays

Local folding free energies, ΔG°_{local}, are calculated using fitted Ising parameters in Table 3.1 (Supporting Materials and Methods). Summing all statistical weights for states where the i^{th} repeat is folded and dividing by the statistical weights where the i^{th} repeat is unfolded gives a local equilibrium constant for folding (K_{local}). The local free energy of folding is calculated using the formula

$$\Delta G_{local} = -RT \ln K_{local} \quad (4)$$

3.4 Results

Consensus TALE (cTALE) design

To design a consensus TALE repeat sequence for Ising analysis, the HMMsearch tool was used to collect and align 3,667 TALE repeat sequences (27). From this alignment, Skylign was used to create an HMM logo (Fig. 3.1A), where the height of each residue at each position is proportional to its conservation (28). The most conserved residue at each position was selected for the consensus-based TALE sequence except at position 30, where arginine was chosen instead of cysteine to simplify folding studies. In addition, the two RVD residues, positions 12 and 13, were varied to generate two common recognition sequences (NS and HD; Fig. 3.1B). HD is the most common RVD sequence, and thus conforms to the consensus sequence. In contrast, NS is less frequent, and provides an interesting point of comparison, both for stability studies and in future work, DNA binding.

We initially built consensus TALE arrays without terminal capping repeats, but found these constructs to self-associate by sedimentation velocity analytical ultracentrifugation (AUC; data not shown). Previous studies with ankyrin repeat and TPR consensus constructs have shown that polar N- and C-terminal caps are essential for solubility (19, 20, 22, 23, 29, 30). Thus, we designed N- and C-terminal caps to help solubilize our consensus TALE arrays.

For the N-terminal cap, we selected a conserved 149 residue N-terminal extension of the naturally occurring TALE gene product PthXo1. Crystal structures show that this N-terminal extension forms four cryptic repeats, with similar structure to TALE repeats despite significant sequence differences from the TALE consensus (Fig. 3.1C) (31, 32). This N-cap has been shown to be resistant to proteolysis and is required for full transcription activation(31). The C-terminal cap was designed by changing consensus hydrophobic residues predicted to be solvent exposed to polar or charged residues (Fig. 3.1D). Sedimentation velocity AUC experiments demonstrated that constructs with both the N- and C-caps are monomeric (Fig. 3.S3 in the Supporting Material). Subsequent AUC experiments showed singly-capped (either N- or C-terminal) constructs are also soluble and monomeric. Including these singly-capped constructs allows us to resolve the thermodynamic contributions of the capping repeats from the internal repeats.

To confirm that cTALE repeats have α-helical secondary structure, far-UV circular dichroism (CD) spectra were collected for various cTALE constructs. The spectra of all constructs are consistent with α-helical structure, and are similar in

shape to the far-UV CD spectrum of the naturally-occurring PthXo1 TALE domain (Fig. 3.2 and 3.S4 in the Supporting Material). Helical structure is retained when either the N- or C- cap is absent. Also, consensus TALEs retain DNA-binding activity (Fig. 3.S5).

The RVD sequence affects cTALE stability

To determine the effect of RVD sequence on stability, urea-induced unfolding transitions were measured by CD spectroscopy for a construct with NS RVDs ($(NS)_6C$, with six NS repeats and a C-cap) and an otherwise identical construct with HD repeats ($(HD)_6C$, Fig. 3.3). Both unfolding transitions are sigmoidal and are well-fitted with a two-state model. The unfolding transitions of the HD and NS constructs have similar slopes (and thus, similar *m*-values), but have significantly different unfolding midpoints (C_m values). As a result, the NS and HD constructs have different free energies of unfolding ($\Delta G^{\circ}_{H_2O}$). The sigmoidal transitions and high *m*-values are consistent with a high level of cooperativity in unfolding, suggesting strong interfacial coupling between repeats. The differences in $\Delta G^{\circ}_{H_2O}$ values indicate that RVD identity affects intrinsic folding energy, inter-repeat coupling energy, or both.

Length and capping dependence of NS TALE stability

To resolve the intrinsic stability from the interfacial coupling energy between TALE repeats, urea-induced unfolding transitions were measured for constructs with different numbers of NS repeats. Because the number of repeats and interfaces in each construct differ, analyzing the unfolding of constructs of different lengths allows the intrinsic and interfacial energies to treated as

independent variables (26). To account for sequence differences in the N and C-terminal repeat sequences, we included constructs lacking either the N- or the C-terminal cap. These constructs are crucial for untangling intrinsic and interfacial free energies from variations due to capping substitutions.

For a given capping structure, adding internal NS repeats increases stability (compare $N(NS)_5C$ and $N(NS)_6C$, as well as $N(NS)_6$, $N(NS)_7$, and $N(NS)_8$; Fig. 3.4A), again consistent with strong energetic coupling between repeats. For constructs with six internal NS repeats, the construct containing both the N- and C-terminal cap has the highest midpoint, followed by the construct with only the N-terminal cap (Fig. 3.4B). The construct with only the C-terminal cap has the lowest midpoint.

Transitions for N-capped constructs are not as steep as for constructs lacking N-caps. This suggests that the unfolding transitions of N-capped constructs are not two-state. It seems unlikely that adding the N-terminal cap uncouples the $(NS)_6C$ unfolding transition; rather, the decreased slope for the N-capped constructs suggests weak coupling between the N-cap and central repeats combined with a high intrinsic stability for the N-term capping segment.

Length and capping dependence of HD TALE stability

To better understand the origins of the stability differences among different RVD sequences (Fig. 3.3), urea-induced unfolding transitions were measured for HD-type repeats of various lengths and capping structures. Adding an internal HD repeat is stabilizing, as shown in Fig. 3.4C. Although the C_m values increase with repeat number for the HD series (compare $(HD)_6C$ and $(HD)_7C$ in Fig. 3.4C),

this increase is smaller than the C_m increase for an NS-type repeat (compare $N(NS)_5C$ and $N(NS)_6C$ in Fig. 3.4A). This suggests that addition of an NS-type repeat is more stabilizing than the addition of an HD-type repeat.

The suggestion that NS-type repeats are more stabilizing than HD-type repeats appears to be at odds with the observation that $(HD)_6C$ is more stable than $(NS)_6C$ (Fig. 3.3). One possible explanation for this apparent discrepancy is that stability differences between NS- and HD-type repeats are unevenly distributed between intrinsic folding and interfacial interaction energies.

Stabilities of TALEs containing mixtures of NS and HD TALE repeats

Naturally occurring TALE proteins contain “mixed RVDs”, meaning that adjacent repeats have different RVD sequences. Fig. 3.S6 in the Supporting Material shows urea-induced unfolding transitions for constructs containing both the NS and HD RVDs, $(HD)_2(NS)_1(HD)_2C$ and $(NS)_1(HD)_5C$. Both mixed RVD constructs have cooperative urea-induced unfolding transitions (Fig. 3.S6), with *m*-values similar to constructs composed solely of one type of repeat. These observations suggest that the size of the cooperative unit is similar for mixed and unmixed constructs.

Global Ising analysis of NS and HD TALE repeat unfolding transitions

To dissect contributions of intrinsic and interfacial stabilities to repeat-protein folding, one-dimensional Ising models were fitted to urea-induced unfolding transitions (19, 20, 22). These models represent unfolding at the level of individual repeats, and account for all combinations of folded and unfolded repeats. The free energy of each of these 2^n configurations (where n is the

number of repeats) is modeled to the sum of the intrinsic energies of each folded repeat and the interfacial interaction energies of pairs of adjacent folded repeats. Because the caps differ in sequence, intrinsic energies are modeled as different from those of the internal, sequence-identical repeats.

Fig. 3.4 and S6 show a global fit of the Ising model to unfolding transitions for NS and HD constructs of different lengths and capping structures in which all NS-, HD-, and mixed RVD constructs are fitted simultaneously. Although each unfolding transition has separate baseline parameters, the entire family of transitions share 11 globally fitted thermodynamic parameters (Table 3.1). To estimate parameter uncertainties, 2000 iterations of bootstrapping were performed, and 95% confidence intervals were calculated (26).

All repeats have unfavorable intrinsic folding free energies and favorable interfacial free energies, consistent with previous Ising analyses of ankyrin repeat and some TPR proteins (19, 20, 22, 23). Intrinsic folding of individual NS repeats is more unfavorable than repeats with the HD RVD. In contrast, adjacent NS repeats are more strongly coupled than adjacent HD repeats. For interfaces between repeats with different RVD sequences, coupling is the same as for the first repeat type in the pair. That is, the identity of the RVD determines the coupling energy to the next repeat (but not the previous repeat).

3.5 Discussion

Analysis of unfolding transitions of TALE constructs of different length and RVD sequence using a nearest-neighbor Ising model provides information on coupling energies, local folding free energies, and RVD sequence-stability

correlation. We find that although changes in the residues responsible for conferring DNA binding specificity have moderate effects on the constructs studied here, these sequence changes have large effects on the distribution of stability within and between repeats, and on the cooperativity of TALE repeat arrays. These results suggest that TALE proteins used for genome editing have different local as well as global stabilities based on the RVDs chosen for DNA sequence recognition.

The identity of the RVD affects stability and cooperativity

Ising analysis of the TALE constructs studied here reveals intrinsically unstable repeat units coupled by stabilizing interfaces, consistent with studies from other linear repeat proteins. This partitioning results in an overall cooperative folding transition, and gives rise to an increase in native-state stability with repeat number. However, the magnitude of the partitioning varies depending on RVD-type. For NS RVDs the intrinsic folding free energy is +5.9 kcal/mol, and the interfacial free energy is -7.8 kcal/mol. For HD RVDs the intrinsic folding energy is +3.5 kcal/mol, and the interfacial energy is -5.0 kcal/mol. That is, individual folded NS-type repeats are more intrinsically unstable, but couple more strongly with their folded neighbors. One result of this difference in partitioning is that NS-type repeats are more strongly coupled than HD-type repeats.

Because naturally occurring TALE proteins are composed of many different RVD sequences, “mixed interfaces” are formed between repeats with different RVDs. Our Ising analysis of mixed NS and HD TALE RVDs shows that

for a pair of mixed RVDs, the interfacial energy is determined by the N-terminal repeat. That is, $\Delta G_{NS,i:HD,i+1} = \Delta G_{NS,i:NS,i+1}$ and $\Delta G_{HD,i:NS,i+1} = \Delta G_{HD,i:HD,i+1}$ (Table 3.1).

Weak coupling of the N-terminal cap

Although the conserved N-cap is required for DNA binding and transcription activation of the naturally occurring PthXo1 TALE array (8, 13, 31, 33–35), we find that this cap only modestly enhances stability of the central repeats. Our Ising analysis is consistent with this observation, showing that the N-cap is intrinsically stable (-4.5 kcal/mol), but it is weakly coupled (-0.82 kcal/mol) to the central repeats. In one proposed mechanism for DNA binding, the N-cap nonspecifically binds DNA and facilitates diffusion (31, 36). Weak coupling of the N-cap from the central repeats could uncouple nonspecific diffusive association from tight sequence-specific DNA binding of central repeats. In such a model, the N-cap acts to increase local concentration of TALEs on DNA while the central repeat domain can separately search for specific sequences.

TALE arrays significantly populate partly folded states

Proteins fold in a highly cooperative manner, rarifying populations of partially folded states. The Ising model allows us to determine the populations of partly folded states. These populations can be represented as a distribution of the local folding free energy of each repeat as a function of position (Fig. 3.5). We define folding free energy as the free energy of all states in which a given repeat is folded minus that of all states where the repeat is unfolded (see

Supporting Materials and Methods). In addition to providing a picture of end-fraying, these ΔG°_{local} distributions provide a picture of accessibility of conformations that are unfolded (i.e., "broken") in the middle of the array.

For short (one to four repeat) homopolymeric TALE arrays (Fig. 3.5A, B), ΔG°_{local} is positive, meaning it is more probable for repeats to be unfolded than folded. For homopolymer constructs with five or more repeats, ΔG°_{local} becomes negative. Central repeats have more negative ΔG°_{local} values than terminal repeats, consistent with end fraying. Somewhat surprisingly, the local folding free energies of internal repeats reach a plateau for long TALE arrays (15 and 20 repeats). The local folding free energies plateau at a lower value for NS-type repeats, that is, central NS-repeats are more stable than central HD-repeats (dashed lines, Fig. 3.5A, B).

The values of local folding free energy plateaus are nearly equal to the sum of a single intrinsic energy and two interfacial energies (since two interfaces must be disrupted in order to fold an internal repeat). The close agreement between these two quantities indicates that the dominant partially unfolded state captured by the Ising model is one in which a single internal repeat unfolds, while the remainder of the TALE array remains folded. This kind of local "break" in the TALE array should disrupt the superhelix and allow direct DNA binding.

Because the intrinsic and interfacial energies for NS and HD-type repeats are different, the plateau values for NS and HD-type repeats are different (-9.7 kcal/mol for NS-type repeats and -6.6 kcal/mol for HD-type repeats, Fig. 3.5A, B).

Thus, the probability of a local unfolding reaction inside of the TALE array depends significantly (by a factor of $10^{3.1/RT}$) on RVD sequence.

Fig. 3.5C shows the calculated ΔG°_{local} distribution for mixed constructs. Alternating NS- and HD-type repeats leads to significant heterogeneity in the ΔG°_{Local} distribution due to the difference in the intrinsic stabilities of NS and HD-type repeats. Compared to unmixed arrays, mixing NS- and HD-type repeats increases the local folding free energy of NS-type repeats while decreasing the local folding free energy of HD-type repeats (Fig. 3.5C). NS repeats have higher ΔG°_{local} values in the mixed repeat array because the N-terminal interface in the mixed system is less favorable. HD repeats have lower ΔG°_{local} values in the mixed repeat array because the N-terminal interface in the mixed system is more favorable.

Natural TALEs, such as PthXo1, contain a diverse set of RVD sequences. This extensive mixing of repeats may result in an increase in the population of partly folded states, either through end fraying, internal repeat unfolding, or interfacial fracture (see below and Fig. 3.S7 in the Supporting Material). The far UV CD spectrum of heteropolymeric PthXo1 has less α-helical signal compared to the far UV CD spectra of homopolymeric cTALEs (Fig. 3.S4 in the Supporting Material). Regions of local instability may be important for facilitating binding to DNA.

TALEs access "fractured" states

A more limited type of local structural distortion is the disruption of a single interface between two folded repeats, which can be viewed as a break internal to the folded TALE array. Such states are not included in the Ising model, which assumes that adjacent folded repeats are automatically coupled through interfacial interaction. To capture these fractures, we modified the Ising partition function to include these states (Supporting Materials and Methods). The probability of fracture is calculated for arrays of 20 cTALE repeats (homopolymeric and mixed arrays, Fig. 3.S7). Mixed cTALEs have the greatest fracture probability, while homopolymeric NS cTALEs have the lowest fracture probability ($3.9x10^{-3}$ versus $1.9x10^{-5}$ respectively). For comparison, the probability of fracture is calculated for consensus Ankyrin repeats (cANKS), another helical 34 residue linear repeat array that has been analyzed using the Ising formalism (19, 22); arrays of cTALEs have fracture probabilities 4-6 orders of magnitude greater than arrays of cANKs.

We have described several different ways cTALEs break: end fraying, unfolding of internal repeats, and rupturing of interfaces. Calculated probabilities for these different types of breakage are compared in Fig. 3.S7. States where terminal repeats are unfolded (end frayed) have the greatest probability. For cTALEs, states where internal repeats are unfolded (internally unfolded) and states where consecutive repeats are folded but uncoupled (interfacially fractured) are significantly populated. These types of structural distortion are energetically accessible to cTALEs and may provide access to a state that is competent for DNA binding, thus increasing the overall binding rate.

Length-dependent stability switch

Because of differences in the energetic partitioning for the NS and HD RVDs, the relative stabilities of these two arrays are predicted to switch as repeats are added. At low repeat number, constructs containing NS-type repeats are less stable than constructs containing the HD RVD. This can be seen in Fig. 3.3, where the midpoint of the $(NS)_6C$ transition is at lower urea concentration than that of $(HD)_6C$.

However, at high repeat number, the Ising model predicts that constructs composed of HD RVDs are less stable than constructs composed of NS RVDs. Although we have been unable to purify these longer constructs ($N(HD)_7$ and $N(HD)_8$ aggregate according to sedimentation velocity), we can use Ising parameters to estimate the free energy of the native state. (Fig. 3.6). At one repeat, the stability is equal to the intrinsic stability; thus, the “native state” of a single HD-repeat construct has a lower free energy than a single NS-repeat construct. Upon the addition of subsequent repeats, native state free energy decreases linearly, with a slope equal to the sum of the intrinsic and interfacial energies. Because this sum is larger (more negative) for NS repeats, the two free energy lines intersect between 7 and 8 repeats.

3.6 Conclusions

Through manipulation of capping sequences, we have designed a consensus TALE array that permits both length and RVD sequence variation, and have found conditions where folded constructs remain monomeric, even with only one terminal cap. Together, our set of constructs satisfies the criteria for

one dimensional Ising analysis, permitting high-precision determination of intrinsic stabilities and nearest-neighbor coupling energies. All constructs show moderate cooperativities (favorable coupling energies, unfavorable intrinsic stabilities). Importantly, these parameters are strongly dependent on RVD sequence. Although we have only looked at NS- and HD-RVDS, it is clear from our studies that RVD sequence identity influences not only global stability, but also intrinsic stability, coupling energy, cooperativity, and accessibility of partly folded states. Together, these factors have implications for genome editing: depending on the sequence of the genomic target, the thermodynamic profile of the cognate TALE may affect the ability to activate target sites. Taking these factors into account in genome editing experiments could improve activity.

Author Contributions

K.G. conducted experiments. K.G. and D.B. designed experiments and wrote the paper.

Acknowledgments

We thank Dr. Barry Stoddard for supplying the PthXo1 plasmid, Dr. Jake Marold for writing the original program used for Ising analysis fitting, Dr. Katherine Tripp and the JHU Center for Molecular Biophysics for instrument access and technical support, Dr. Evangelos Moudrianakis for assistance with analytical ultracentrifugation, and Dr. Michael McCaffery and the JHU Integrated Imaging Center for technical support. This work was supported by NIH grant R01 GM068462 to D.B. K.R.G. was supported by NIH training grant T32-GM008403.

References

1. Kay, S., S. Hahn, E. Marois, G. Hause, and U. Bonas. 2007. A bacterial effector acts as a plant transcription factor and induces a cell size regulator. Science. 318: 648–651.

2. Römer, P., S. Hahn, T. Jordan, T. Strauss, U. Bonas, and T. Lahaye. 2007. Plant pathogen recognition mediated by promoter activation of the pepper Bs3 resistance gene. Science. 318: 645–648.

3. Boch, J., and U. Bonas. 2010. Xanthomonas AvrBs3 family-type III effectors: discovery and function. Annu. Rev. Phytopathol. 48: 419–436.

4. Boch, J., H. Scholze, S. Schornack, A. Landgraf, S. Hahn, S. Kay, T. Lahaye, A. Nickstadt, and U. Bonas. 2009. Breaking the code of DNA binding specificity of TAL-type III effectors. Science. 326: 1509–1512.

5. Moscou, M.J., and A.J. Bogdanove. 2009. A simple cipher governs DNA recognition by TAL effectors. Science. 326: 1501.

6. Miller, J.C., L. Zhang, D.F. Xia, J.J. Campo, I.V. Ankoudinova, D.Y. Guschin, J.E. Babiarz, X. Meng, S.J. Hinkley, S.C. Lam, D.E. Paschon, A.I. Vincent, G.P. Dulay, K.A. Barlow, D.A. Shivak, E. Leung, J.D. Kim, R. Amora, F.D. Urnov, P.D. Gregory, and E.J. Rebar. 2015. Improved specificity of TALE-based genome editing using an expanded RVD repertoire. Nat. Methods. 12: 465–471.

7. Li, T., S. Huang, W.Z. Jiang, D. Wright, M.H. Spalding, D.P. Weeks, and B. Yang. 2011. TAL nucleases (TALNs): hybrid proteins composed of TAL effectors and FokI DNA-cleavage domain. Nucleic Acids Res. 39: 359–372.

8. Christian, M., T. Cermak, E.L. Doyle, C. Schmidt, F. Zhang, A. Hummel, A.J. Bogdanove, and D.F. Voytas. 2010. Targeting DNA double-strand breaks with TAL effector nucleases. Genetics. 186: 757–761.

9. Cong, L., R. Zhou, Y.-C. Kuo, M. Cunniff, and F. Zhang. 2012. Comprehensive interrogation of natural TALE DNA-binding modules and transcriptional repressor domains. Nat. Commun. 3: 968.

10. Geissler, R., H. Scholze, S. Hahn, J. Streubel, U. Bonas, S.-E. Behrens, and J. Boch. 2011. Transcriptional activators of human genes with programmable DNA-specificity. PloS One. 6: e19509.

11. Li, Y., R. Moore, M. Guinn, and L. Bleris. 2012. Transcription activator-like effector hybrids for conditional control and rewiring of chromosomal transgene expression. Sci. Rep. 2: 897.

12. Mahfouz, M.M., L. Li, M. Piatek, X. Fang, H. Mansour, D.K. Bangarusamy, and J.-K. Zhu. 2012. Targeted transcriptional repression using a chimeric TALE-SRDX repressor protein. Plant Mol. Biol. 78: 311–321.

13. Zhang, F., L. Cong, S. Lodato, S. Kosuri, G.M. Church, and P. Arlotta. 2011. Efficient construction of sequence-specific TAL effectors for modulating mammalian transcription. Nat. Biotechnol. 29: 149–153.

14. Maeder, M.L., J.F. Angstman, M.E. Richardson, S.J. Linder, V.M. Cascio, S.Q. Tsai, Q.H. Ho, J.D. Sander, D. Reyon, B.E. Bernstein, J.F. Costello, M.F. Wilkinson, and J.K. Joung. 2013. Targeted DNA demethylation and activation of endogenous genes using programmable TALE-TET1 fusion proteins. Nat. Biotechnol. 31: 1137–1142.

15. Miyanari, Y., C. Ziegler-Birling, and M.-E. Torres-Padilla. 2013. Live visualization of chromatin dynamics with fluorescent TALEs. Nat. Struct. Mol. Biol. 20: 1321–1324.

16. Ma, H., P. Reyes-Gutierrez, and T. Pederson. 2013. Visualization of repetitive DNA sequences in human chromosomes with transcription activator-like effectors. Proc. Natl. Acad. Sci. U. S. A. 110: 21048–21053.

17. Kim, H., and J.-S. Kim. 2014. A guide to genome engineering with programmable nucleases. Nat. Rev. Genet. 15: 321–334.

18. Deng, D., C. Yan, X. Pan, M. Mahfouz, J. Wang, J.-K. Zhu, Y. Shi, and N. Yan. 2012. Structural basis for sequence-specific recognition of DNA by TAL effectors. Science. 335: 720–723.

19. Aksel, T., A. Majumdar, and D. Barrick. 2011. The contribution of entropy, enthalpy, and hydrophobic desolvation to cooperativity in repeat-protein folding. Struct. Lond. Engl. 1993. 19: 349–360.

20. Kajander, T., A.L. Cortajarena, E.R.G. Main, S.G.J. Mochrie, and L. Regan. 2005. A new folding paradigm for repeat proteins. J. Am. Chem. Soc. 127: 10188–10190.

21. Mello, C.C., and D. Barrick. 2004. An experimentally determined protein folding energy landscape. Proc. Natl. Acad. Sci. U. S. A. 101: 14102–14107.

22. Wetzel, S.K., G. Settanni, M. Kenig, H.K. Binz, and A. Plückthun. 2008. Folding and unfolding mechanism of highly stable full-consensus ankyrin repeat proteins. J. Mol. Biol. 376: 241–257.

23. Marold, J.D., J.M. Kavran, G.D. Bowman, and D. Barrick. 2015. A Naturally Occurring Repeat Protein with High Internal Sequence Identity Defines a New Class of TPR-like Proteins. Struct. Lond. Engl. 1993. .

24. Cermak, T., E.L. Doyle, M. Christian, L. Wang, Y. Zhang, C. Schmidt, J.A. Baller, N.V. Somia, A.J. Bogdanove, and D.F. Voytas. 2011. Efficient design and assembly of custom TALEN and other TAL effector-based constructs for DNA targeting. Nucleic Acids Res. 39: e82.

25. Poland, D., and H.A. Scheraga. 1970. Theory of helix-coil transitions in biopolymers: statistical mechanical theory of order-disorder transitions in biological macromolecules. Academic Press.

26. Aksel, T., and D. Barrick. 2009. Analysis of repeat-protein folding using nearest-neighbor statistical mechanical models. Methods Enzymol. 455: 95–125.

27. Finn, R.D., J. Clements, and S.R. Eddy. 2011. HMMER web server: interactive sequence similarity searching. Nucleic Acids Res. 39: W29–37.

28. Wheeler, T.J., J. Clements, and R.D. Finn. 2014. Skylign: a tool for creating informative, interactive logos representing sequence alignments and profile hidden Markov models. BMC Bioinformatics. 15: 7.

29. Mosavi, L.K., and Z.-Y. Peng. 2003. Structure-based substitutions for increased solubility of a designed protein. Protein Eng. 16: 739–745.

30. Tripp, K.W., and D. Barrick. 2007. Enhancing the stability and folding rate of a repeat protein through the addition of consensus repeats. J. Mol. Biol. 365: 1187–1200.

31. Gao, H., X. Wu, J. Chai, and Z. Han. 2012. Crystal structure of a TALE protein reveals an extended N-terminal DNA binding region. Cell Res. 22: 1716–1720.

32. Mak, A.N.-S., P. Bradley, R.A. Cernadas, A.J. Bogdanove, and B.L. Stoddard. 2012. The crystal structure of TAL effector PthXo1 bound to its DNA target. Science. 335: 716–719.

33. Miller, J.C., S. Tan, G. Qiao, K.A. Barlow, J. Wang, D.F. Xia, X. Meng, D.E. Paschon, E. Leung, S.J. Hinkley, G.P. Dulay, K.L. Hua, I. Ankoudinova, G.J. Cost, F.D. Urnov, H.S. Zhang, M.C. Holmes, L. Zhang, P.D. Gregory, and E.J. Rebar. 2011. A TALE nuclease architecture for efficient genome editing. Nat. Biotechnol. 29: 143–148.

34. Mussolino, C., R. Morbitzer, F. Lütge, N. Dannemann, T. Lahaye, and T. Cathomen. 2011. A novel TALE nuclease scaffold enables high genome editing activity in combination with low toxicity. Nucleic Acids Res. 39: 9283–9293.

35. Sun, N., J. Liang, Z. Abil, and H. Zhao. 2012. Optimized TAL effector nucleases (TALENs) for use in treatment of sickle cell disease. Mol. Biosyst. 8: 1255–1263.

36. Meckler, J.F., M.S. Bhakta, M.-S. Kim, R. Ovadia, C.H. Habrian, A. Zykovich, A. Yu, S.H. Lockwood, R. Morbitzer, J. Elsäesser, T. Lahaye, D.J. Segal, and E.P. Baldwin. 2013. Quantitative analysis of TALE-DNA interactions suggests polarity effects. Nucleic Acids Res. 41: 4118–4128.

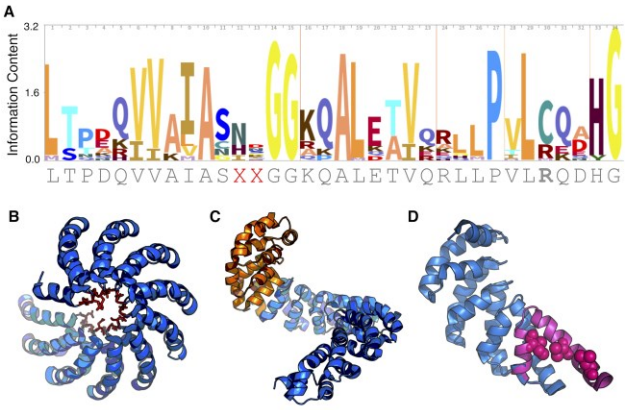


Figure 3.1. Sequence conservation and structure of TALE repeats. (A) A sequence logo of TALE repeats showing conservation at each of the 34 positions. The sequence below is the consensus sequence used in these studies. At the RVD positions (12 and 13), two common recognition motifs were selected (NS and HD). (B) Crystal structure of dHax3 highlighting location of RVDs (red sticks)(18). (C) The N-terminal cap (orange) is a conserved extension of the repeat domain composed of four TALE-like repeats(31). (D) The C-terminal cap was designed by substituting solvent exposed hydrophobic residues to polar or charged residues (pink spheres).

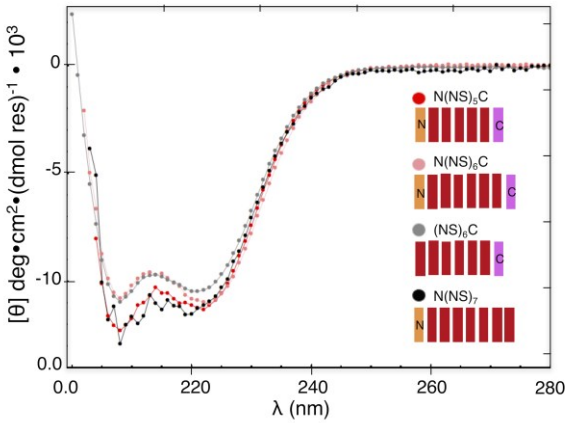


Figure 3.2. Doubly- and singly-capped TALE consensus constructs are α-helical. Far-UV CD spectra of TALE NS repeat constructs with N-cap, C-cap, and both caps. Spectra are consistent with folded, α-helical structures. Conditions: 300 mM NaCl, 10 mM Tris HCl pH 7.4, 5% glycerol, 25°C.

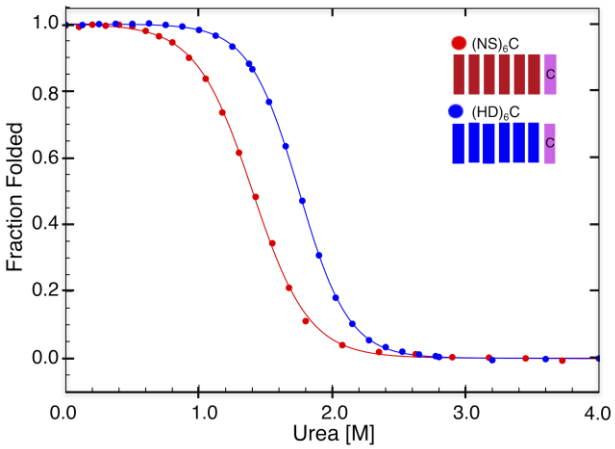


Figure 3.3. Consensus TALE stability is dependent on RVD sequence. Urea-induced unfolding transitions of TALE constructs show cooperative, two-state unfolding transitions. Urea-induced unfolding transitions of each construct were fitted with a two-state model for unfolding (solid lines). Global stabilities, based on unfolding midpoints, vary significantly with RVD sequence, although slopes of the transitions do not. Conditions: 300 mM NaCl, 10 mM Tris HCl pH 7.4, 5% glycerol, 25°C.

Figure 3.4. Length- and capping-dependence of TALE HD- and NS-RVD stability. Urea-induced unfolding transitions of TALE constructs were globally fitted using a heterogeneous nearest-neighbor Ising model (N-capped, dashed lines; C-capped, solid lines; doubly-capped, dotted lines). (A) NS-type repeat constructs with increasing repeat number, for N-capped and doubly-capped constructs. Stability increases with number of repeats. (B) Constructs with six NS-type repeats with varying capping identities. $N(NS)_6C$ is most stable followed by $N(NS)_6$. $(NS)_6C$ has the smallest C_m but the largest slope. (C) HD-type repeat constructs with increasing repeat number. The increase in midpoint of the transition between $(HD)_6C$ and $(HD)_7C$ is less than the increase in midpoint between $N(NS)_5C$ and $N(NS)_6C$ (panel A). (D) Constructs with six HD-type repeats with varying capping identities. As with the NS-type repeats (B), the doubly capped construct is more stable than the singly capped construct. Conditions: 300 mM NaCl, 10 mM Tris HCl pH 7.4, 5% glycerol, 25°C.

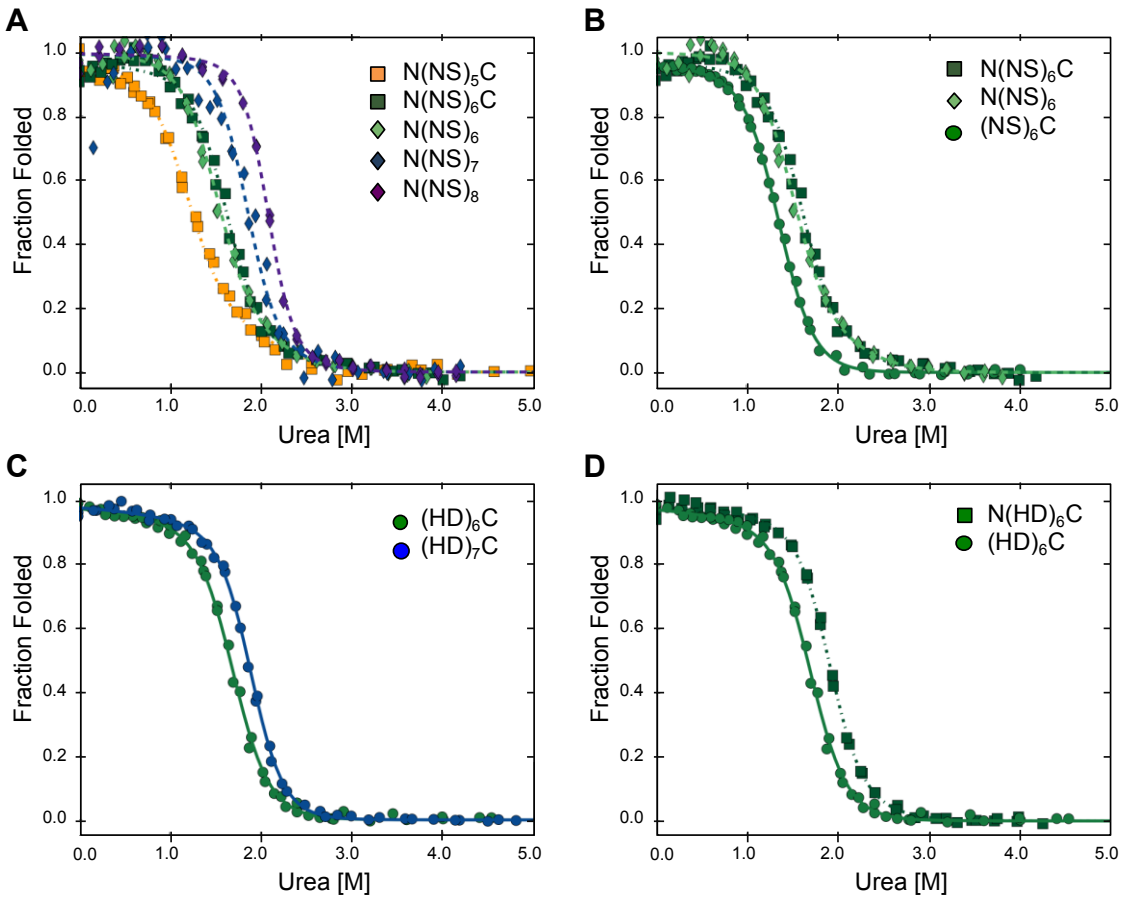
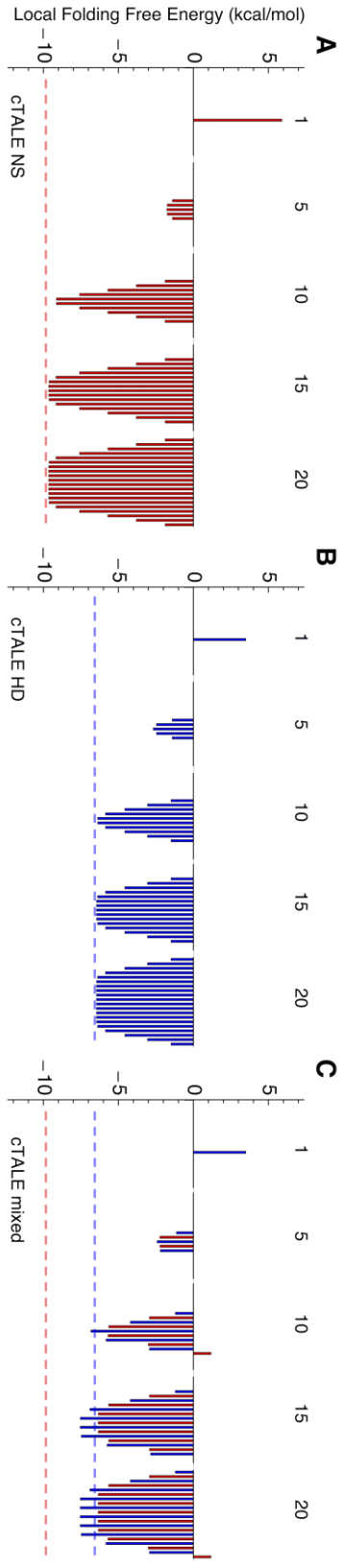

A
Fraction Folded
Urea [M]
N(NS)5C
N(NS)6C
N(NS)6
N(NS)7
N(NS)8
B
Fraction Folded
Urea [M]
N(NS)6C
N(NS)6
(NS)6C
C
Fraction Folded
Urea [M]
(HD)6C
(HD)7C
D
Fraction Folded
Urea [M]
N(HD)6C
(HD)6C

Figure 3.5. Distribution of local folding free energies for TALE repeat arrays as a function of length. At each position, probabilities of a repeat being folded (and unfolded) were calculated from the nearest-neighbor partition function using free energies from the Ising fit. Local folding free energies (ΔG°_{local}) were calculated from Eq. 4. For homopolymeric TALE arrays (A, NS-type repeats in red, B, HD-type repeats in blue) the ΔG°_{local} values of the central repeats decreases with repeat number until a plateau in local folding free energy is reached (shown as dashed lines). This plateau in ΔG°_{local} is lower (i.e., greater stability) for NS repeats (A) than HD repeats (B). Constructs containing both NS and HD repeat types have a heterogeneous stability distribution, with local unfolding of the C-terminal NS repeats (C). Mixing NS (red) and HD (blue) repeats in an alternating fashion decreases ΔG°_{local} of HD repeats, but increases ΔG°_{local} of NS repeats. All constructs show fraying of end repeats, as would be expected from a cooperative nearest-neighbor model.


A
B
C
Local Folding Free Energy (kcal/mol)
5
0
−5
−10
1
5
10
15
20
cTALE NS
cTALE HD
cTALE mixed

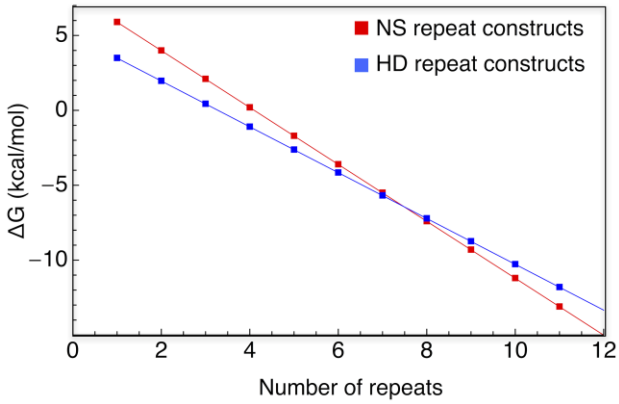


Figure 3.6. Differences in energetic partitioning for NS- and HD-RVDs results in a length dependent stability switch. Using free energies from the Ising fit, folding free energies for arrays with increasing repeat number were calculated. At low repeat number, the fully folded state of NS-type constructs are less stable than HD-type constructs. However, at high repeat number, NS-type constructs are more stable than HD-type constructs. The crossover point is between seven and eight repeats.

Table 3.1 Summary of thermodynamic parameters obtained from Ising-fit.

Intrinsic terms	**ΔG_N**	**ΔG_{NS}**	**ΔG_{HD}**	**ΔG_C**	**m_i**	
Best-fit	-4.42	5.89	3.49	7.14	-0.50	
95% CI[a]	-4.80, -4.03	5.47, 6.33	3.13, 3.82	6.78, 7.50	-0.52, -0.48	
Interfacial terms	**$\Delta G_{N, i+1}$**	**$\Delta G_{NS, NS+1}$**	**$\Delta G_{HD, HD+1}$**	**$\Delta G_{C, HD-1}$**	**$\Delta G_{NS, HD+1}$**	**$\Delta G_{HD, NS+1}$**
Best-fit	-0.85	-7.79	-5.02	-8.49	-7.58	-4.73
95% CI[a]	-0.94, -0.76	-8.30, -7.30	-5.42, -4.59	-9.00, -8.00	-8.09, -7.09	-5.24, -4.19

All values obtained are kcal/mol with the exception of m_i, which is kcal/mol/M. [a]95% confidence intervals are from 2000 iterations of bootstrap analysis.

3.7.1 Supporting Figures

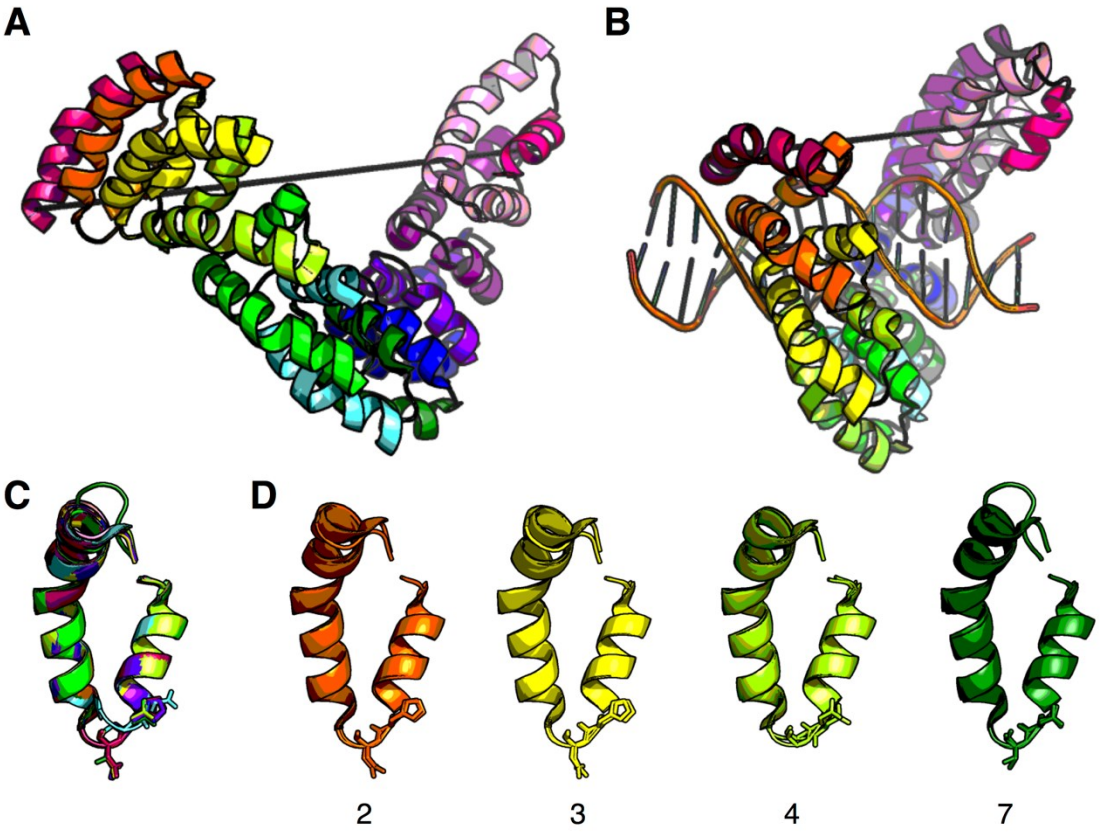


Figure 3.S1. The TALE conformational change upon DNA binding is mediated by small changes propagated through many repeats. (A) DNA-free structure of dHax3 (colored by repeat, PDB:3V6P). The distance between C_α of residue 304 to C_α of residue 666 is 74 Å. (B) DNA-bound structure of dHax3 (PDB:3V6T). The distance between C_α of residue 304 to C_α of residue 666 is 50 Å. (C) Alignment of C_α s of all 11 repeats in the DNA-bound structure. RMSDs range from 0.3 to 0.6 $Å^2$. (D) Alignment of C_α s of repeats 2, 3, 4, and 7 in the DNA-free and DNA-bound structure. RMSDs range from 0.3 to 0.6 $Å^2$.

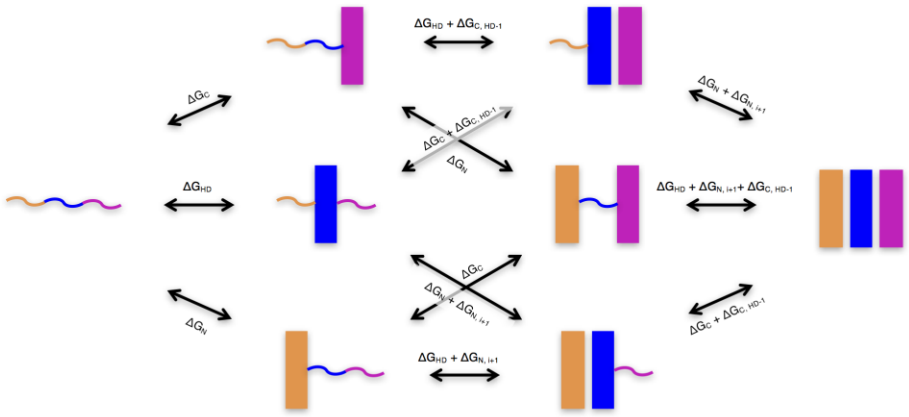


Figure 3.S2. A 1-D Ising Model for a cTALE with three repeats. In this example, a construct with the one N-cap, one HD repeat, and one C-cap is shown. The partition function *q* is used to derive the dependence of the fraction folded repeats (f_{folded}) on urea concentration (See equations 1-4). The folding of each repeat results in an intrinsic free energy change, along with free energy changes from formation of one or two interfaces (for terminal repeat unfolding and internal repeat folding respectively).

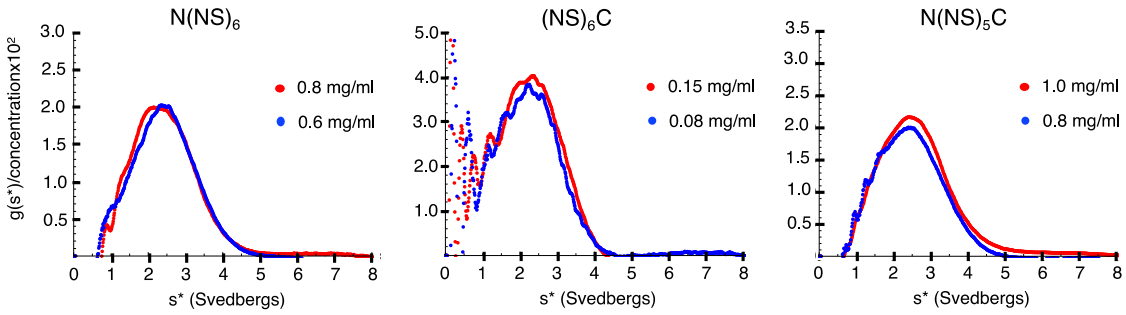


Figure 3.S3. Sedimentation Velocity g(s*) plots for capped consensus TALE repeats. Sedimentation velocity experiments were performed with $N(NS)_6$, $(NS)_6C$, and $N(NS)_5C$, as described in Supporting Materials and Methods. The g(s*) distributions are consistent with monomers. Conditions: 300 mM NaCl, 10 mM Tris HCl pH 7.4, 5% glycerol, 25°C.

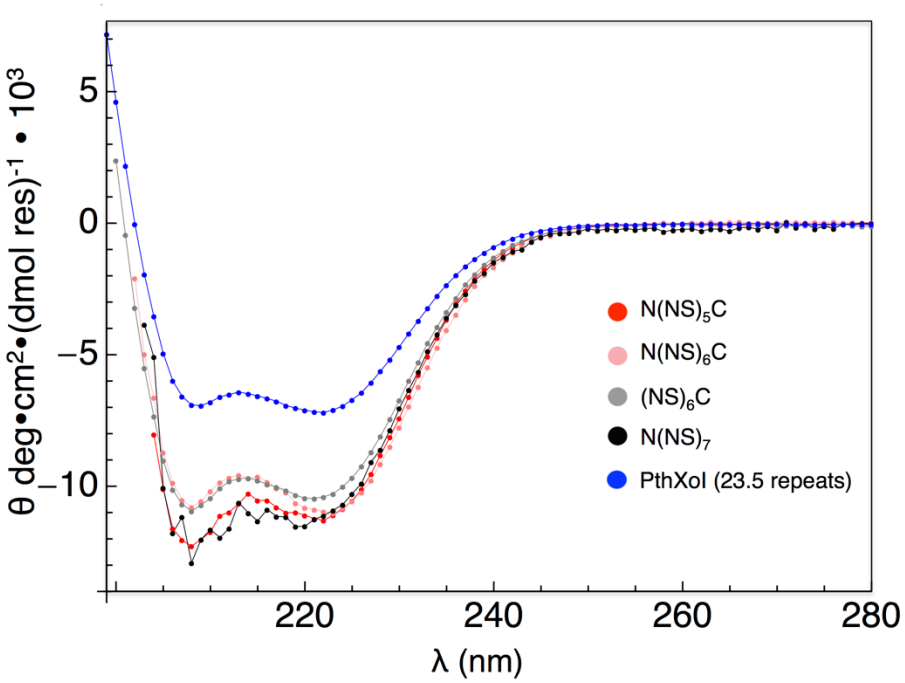


Figure 3.S4. Consensus and natural TALEs and have α-helical secondary structure. FarUV CD of cTALEs and the repeat domain of PthXo1, a naturally-occurring TALE, show similar shape. PthXo1 has less helical structure than the cTALEs.

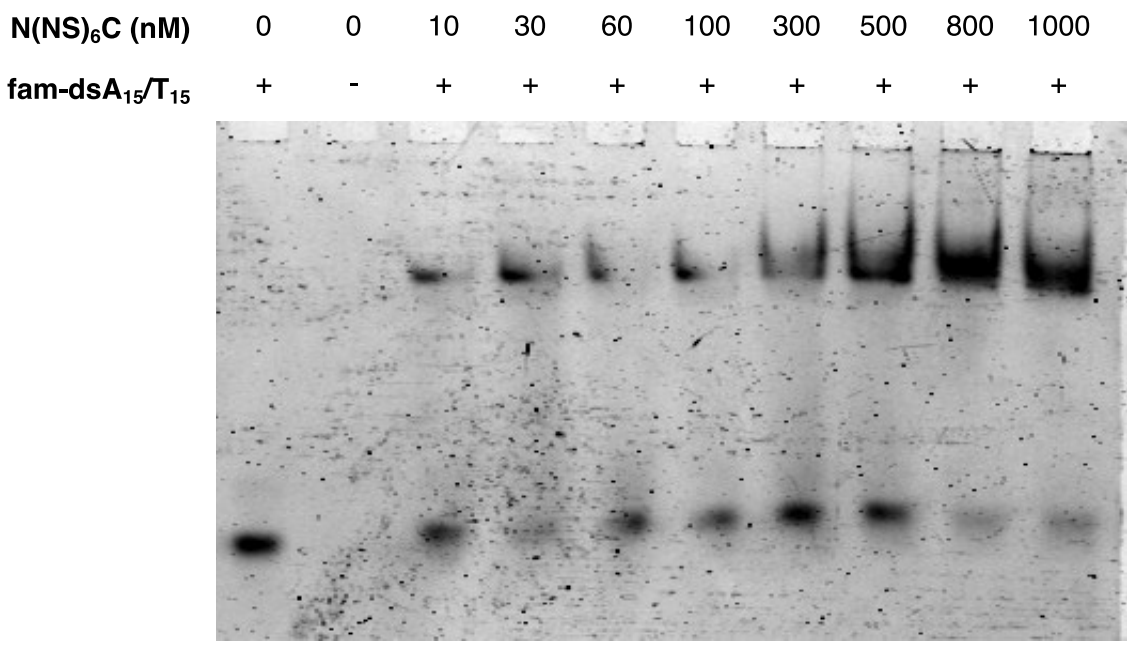


Figure 3.S5. Consensus TALEs bind double stranded DNA. An electrophoretic mobility shift assay (EMSA) shows $N(NS)_6C$ binding 10 nM fam-dsA_{15}/T_{15} (see Supporting Materials and Methods). Conditions: 150 mM KCl, 0.1 mM DTT, 10 mM Tris, pH 7.4, 25% sucrose.

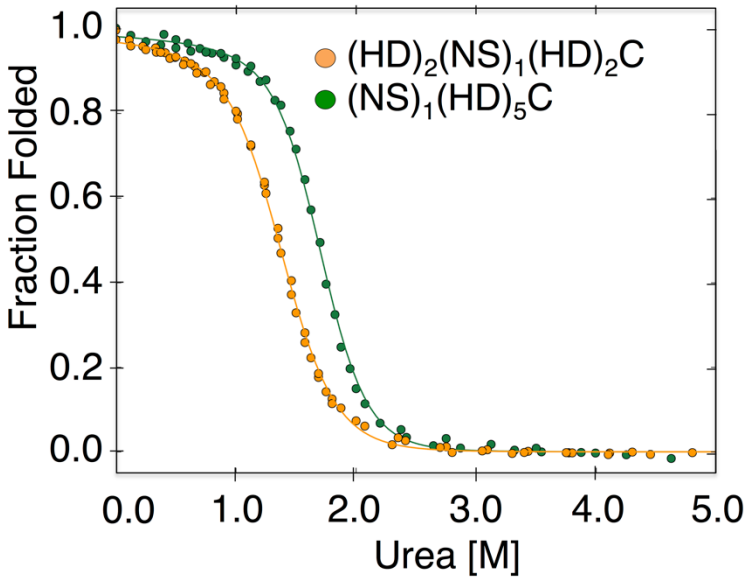


Figure 3.S6. "Mixed RVD" constructs have single cooperative unfolding transition. Urea-induced unfolding transitions for constructs containing both NS- and HD-type repeats. Transitions are cooperative and slopes of $(HD)_2(NS)_1(HD)_2C$ and $(NS)_1(HD)_5C$ are similar to $(NS)_6C$, suggesting cooperative unfolding units of similar size.

Figure 3.S7. Calculated probabilities for fully folded and broken TALEs. Probabilities of fully folded, end frayed, internally unfolded, and interfacially fractured populations were calculated using free energies from Ising analysis of cTALEs and, for comparison, consensus ankyrin repeat proteins (cANK)(19). Calculations were performed on arrays of 20 repeats. Microstates containing unfolded repeats in the regions 1-5 or 16-20 were included in the calculation of $p_{\text{end-frayed}}$. Microstates containing unfolding in repeats 6-15 were included in the calculation of $p_{\text{internally unfolded}}$. To calculate $p_{\text{interfacially fractured}}$, a separate Ising-like model was generated that includes microstates with structural deformation between two consecutive uncoupled but folded repeats (Supporting Materials and Methods). There is a significant population of broken states for cTALEs, in contrast to the very low probabilities of broken states calculated for cANKs.

	$p_{fully\ folded}$	$p_{end\ frayed}$	$p_{internally\ unfolded}$	$p_{interfacially\ fractured}$
cTALE (NS)	0.92	0.08	8.3×10^{-7}	1.9×10^{-5}
cTALE(HD)	0.85	0.15	1.7×10^{-4}	3.9×10^{-3}
cTALE(mix)	0.11	0.89	1.1×10^{-4}	3.1×10^{-3}
cANK	>0.999	<0.001	1.5×10^{-14}	8.4×10^{-9}

BROKEN

fully folded | end frayed | internally unfolded | interfacially fractured

3.7.2 Supporting Materials and Methods

Sedimentation Velocity

Analytical ultracentrifugation sedimentation velocity experiments were performed using a Beckman XL-1 analytical ultracentrifuge as previously described(22). Proteins were dialyzed in reference buffer for 24 hours prior to centrifugation. Concentrations ranged from 2 to 30 μM. AUC data were analyzed with SEDANAL. DCDT analysis was performed on scans spanning a one hour interval, generating s* value distributions (g(s*) versus s*)(23).

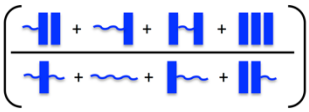
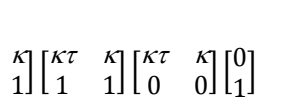
Electrophoretic Mobility Shift Assay (EMSA)

Flourescein (Fam)-labeled A_{15} single stranded DNA was annealed with unlabeled T_{15} single stranded DNA to prepare fam-dsA_{15}/T_{15}. Reactions with increasing concentrations of cTALEs were incubated at room temperature for 20 minutes. Samples were then loaded onto fresh 6% non-denaturing 0.5X TBE gels and were electrophoresed at 100 volts in the cold room for 60 minutes. Gels were imaged on the JHU Integrated Imaging Center Typhoon 9410 Variable Mode Imager and analyzed with ImageJ.

Calculating local folding free energy from intrinsic and interfacial stabilities

Local folding free energies (ΔG°_{local}) describe the free energy difference between all states where a particular repeat, *i*, is folded, and all states where the ith repeat is unfolded (Equation S1). For example, to calculate the ΔG°_{local} of the

third repeat in a three repeat array, partition functions for each construct are created as illustrated in Figure 3.S5. Next, q_3 sums all statistical weights of all conformations where the third repeat is folded. Placing zeros in the third matrix is equivalent to not counting statistical weights when the third repeat is unfolded (Equation S2). Dividing q_3 by the partition function, q (Equation S3), gives the probability that the third repeat is folded (Equation S4). The probability that the third repeat is unfolded is simply one subtracted by the probability of the third repeat being folded. The ΔG°_{Local} of a repeat is the log of the ratio of these probabilities (Equation S5).

$$Local\ Stability = -RT \ln \left(\frac{\text{  }}{\text{  }} \right) \quad \text{(S1)}$$

$$q_3 = \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} \kappa\tau & \kappa \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \kappa\tau & \kappa \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \kappa\tau & \kappa \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} \quad \text{(S2)}$$

$$q = \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} \kappa\tau & \kappa \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \kappa\tau & \kappa \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \kappa\tau & \kappa \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} \quad \text{(S3)}$$

$$p_{3,N} = \frac{q_3}{q} \quad \text{(S4)}$$

$$\Delta G^{\circ}_{Local} = -RT \ln \left(\frac{p_{3,N}}{p_{3,D}} \right) \quad \text{(S5)}$$

An extended Ising model to include fractured states

The 1-D Ising model we used to describe folding can be extended to include states where consecutive repeats are folded but uncoupled. The physical interpretation of such states is that a structural deformation occurs leaving repeats folded but preventing them from forming favorable interactions required for coupling. Such states should be accessible by opening a turn between

helices, requiring changes to just a few backbone dihedral angles. Generation of partition functions for the folding Ising model is described in Equations 1-4 in the Materials and Methods. Equation 2 can be modified to include fractured states by replacing $\kappa\tau$ with $\kappa\tau + \kappa$ in the first column and row of the 2x2 matrices for each repeat.

$$\theta = \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} \kappa_N \tau_N + \kappa_N & \kappa_N \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \kappa_R \tau_R + \kappa_R & \kappa_R \\ 1 & 1 \end{bmatrix}^{n-2} \begin{bmatrix} \kappa_C \tau_C + \kappa_C & \kappa_C \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} \quad \text{(S6)}$$

If the *i* and *i* + 1 repeats can be folded or unfolded, there are four possible states (FF, FU, UF, and UU). Each of the four elements in the 2x2 matrices represent one of these states. Row one, column one in each matrix represents states where the *i* and *i* + 1 repeats are folded (FF). In the folding Ising model, if *i* and *i* + 1 are folded, the statistical weight is $\kappa\tau$ because the interfacial energy is automatically associated with the folding of the i^{th} repeat (through τ). To allow for states where there is no coupling between consecutive repeats, we simply add a term κ to the position where both repeats are folded. This is saying that when the *i* and *i* + 1 repeats are folded (FF), either there is a coupling ($\kappa\tau$) or there is not (κ). Because these two options are mutually exclusive, their contributions to statistical weight sum.

Urea induced unfolding transitions fitted with this extended model returned best-fit values within 0.03% of the best-fit values in Table 3.1, with the exception of the interfacial energy of the N-cap and next repeat ($\Delta G_{N,\,N+1}$). Because the interface between the N-cap and the adjacent repeat is rather weak, the sum $\kappa\tau + \kappa$ in the modified Ising model is significantly larger than the Ising value $\kappa\tau$. As a

result, the best-fit value for $\Delta G_{N, N+1}$ is -0.69 kcal/mol with the fracture Ising model as compared to -0.85 kcal/mol with the folding Ising model.

3.7.3 Supporting References

22. Marold, J.D., J.M. Kavran, G.D. Bowman, and D. Barrick. 2015. A Naturally Occurring Repeat Protein with High Internal Sequence Identity Defines a New Class of TPR-like Proteins. Struct. Lond. Engl. 1993. .

23. Stafford, W.F. 1992. Boundary analysis in sedimentation transport experiments: a procedure for obtaining sedimentation coefficient distributions using the time derivative of the concentration profile. Anal. Biochem. 203: 295–301.

CHAPTER 4

Transcription activator-like effector (TALE) conformational heterogeneity slows observed DNA binding and unbinding.

This chapter is a collaboration with Taekjip Ha and Jaba Mitra at Johns Hopkins University.

4.1 Abstract

Transcription activator-like effectors (TALEs) bind DNA through a domain of tandem 34-residue repeats interacting with DNA one repeat per base pair. In this study, we use single molecule total internal reflection microscopy to examine the kinetic mechanism of DNA binding and unbinding. We see two distinct types of dynamic behavior in the bound state. Using deterministic simulations to analyze the data, we find evidence for conformational heterogeneity in both the free- and DNA-bound TALE arrays. We connect these results with previous work demonstrating populations of partly folded TALE states. In the DNA-bound state, conformational exchange involves transitions from short-lived encounter complexes to longer-lived locked complexes. In the DNA free state, we find the effects of conformational heterogeneity on binding to depend on the length of the TALE array. TALEs that form less than one superhelical turn around DNA (eight repeats) access partly folded open states that inhibit DNA binding, whereas TALEs that form more than one superhelical turn around DNA (sixteen repeats) access partly folded open states that facilitate for DNA binding.

4.2 Introduction

Transcription activator-like effectors (TALEs) are bacterial proteins containing a domain of tandem DNA-binding repeats as well as a eukaryotic transcriptional activation domain[1,2]. The repeat domain binds double stranded DNA with a register of one repeat per base pair, and specificity is determined by the sequence identity at positions twelve and thirteen in each TALE repeat, which are referred to as repeat variable diresidues (RVDs)[3–5]. This specificity code has enabled design of TALE-based tools for transcriptional control[6–10], DNA modifications[11], in-cell microscopy[12,13], and genome editing (TALENs)[14,15].

TALE repeat domains wrap around DNA in a continuous superhelix of 11.5 TALE repeats per turn[16,17]. Because TALEs contain on average 17.5 repeats[18], most form over 1.5 full turns around DNA. Many proteins that form rings around DNA require energy in the form of ATP to open or close around DNA, yet TALEs are capable of wrapping around DNA without energy from nucleotide triphosphate hydrolysis. One possibility is that TALEs bind DNA through an energetically accessible open conformation. Consistent with this possibility, we previously demonstrated that TALE arrays can populate partly folded or broken states that may be more active for DNA binding[19].

Consensus TALEs (cTALEs) are homopolymeric arrays composed of the most commonly observed residue at each of the 34 positions of the repeat[19]. In addition to simplifying analysis of folding and conformational heterogeneity, the consensus approach simplifies analysis of DNA binding, eliminating contributions from sequence heterogeneity and providing an easy means of site-specific

labeling. Whereas the consensus sequence remains constant among all repeats in a cTALE array, two RVDs are examined (NS and HD). While NS-containing repeats are predicted to have high affinity and low base specificity, HD-containing repeats are predicted to have low affinity and high base specificity for cytosine.

Here we characterize DNA binding kinetics of cTALEs using total internal reflection fluorescence single-molecule microscopy. We find that consensus TALE arrays bind to DNA reversibly, with high affinity. Analysis of the dwell-times of the on- and off-states reveals multiphasic binding and unbinding kinetics, suggesting conformational heterogeneity in both the free and DNA bound state. Deterministic simulations support such a model, and provide rate constants for both conformational changes and binding. Comparing the dynamics observed here to previously characterized local unfolding suggests that locally unfolded states inhibit binding of short cTALE arrays (less than one full superhelical turn around DNA), whereas they promote binding of long arrays (more than 1 full superhelical turn).

4.3 Results

cTALE design

Consensus TALE (cTALE) repeat sequence was design previously described[19]. To avoid self- association of cTALE arrays, we fused to a conserved N-terminal extension of the PthXo1 gene. While the sequence of this domain differs from a TALE repeat sequence, this domain forms four TALE repeat

structures when not bound to DNA[17,20] and is required for full transcriptional activation[20]. In this study, all repeat arrays contain the solubilizing N-terminal domain.

cTALEs local instability promotes population of partly folded states

Figure 4.1A depicts different types of partly folded states of a generic repeat protein. In the fully folded state, all repeats are folded, and all interfaces are intact. In the end-frayed state, at least one terminal repeat is unfolded and all interfaces, except the interface(s) between the unfolded and adjacent folded repeat(s), are intact. In the internally unfolded state, one central repeat is unfolded and all interfaces, except the interfaces involving the unfolded repeats, are intact. In the interfacially ruptured state, all repeats are folded and one interface is not intact due to local structural distortion.

Figure 4.1B shows calculated free energies between types of partly folded states and the fully folded repeat array. The distribution of partly folded states is calculated for different 20-repeat arrays containing two types of TALE arrays (with the NS RVD in red and with the HD RVD in blue) as well as consensus ankyrin arrays (cAnk in black). For cTALE arrays, end frayed states are within a few *kT*, internally unfolded states are highest in energy, and interfacially ruptured states fall energetically between end frayed and internally unfolded states. Changing the RVD affects the distribution of these partly folded states: HD repeat containing arrays are more likely to internally unfold or interfacially rupture than NS repeat containing arrays. cTALEs are more likely to populate many of these

partly folded states than cAnk is to populate even the lowest energy partly folded state, end frayed. Thus, compared to ankyrin repeats, cTALEs are locally unstable, meaning they are likely to form partly folded states. As these states disrupt the superhelix, they may facilitate DNA binding.

Single-molecule studies of cTALE binding to DNA

Figure 4.2A shows a schematic of the single-molecule total internal reflection fluorescence (smTIRF) experiments performed to measure DNA binding. For site-specific cTALE labeling, R30 is mutated to cysteine only in the indicated repeat. This position is commonly cysteine in naturally occurring TALEs (in earlier folding studies, arginine was chosen in the consensus sequence to simplify folding studies)[19] This cysteine was Cy3-labelled using maleimide chemistry, and was attached to biotinylated slides via the C-terminal His_6 tag and α-Penta•His antibodies. At salt concentrations below 300 mM NaCl, cTALEs aggregate. Because DNA binding is weak at high salt concentrations, measuring binding kinetics in bulk at high salt is not possible. Tethering cTALEs to the quartz slide prevents self-association, even in the low salt concentrations required to study DNA binding kinetics. Figure 4.2B shows a single-molecule FRET histogram generated from a sample with tethered cTALE (8 NS-type repeats and the N-terminal domain labeled via a cysteine in the first repeat; $N(NS)_8$) and no DNA showing the 0.0 FRET donor-only peak.

To test for DNA binding to tethered cTALE constructs, we flushed with Cy5-labeled DNA ($Cy5.A_{15}/T_{15}$) into channels containing tethered $N(NS)_8$. This

results in a new peak a FRET value of 0.45, indicating that DNA binds directly to TALE arrays. As DNA concentration in the bulk phase is increased, the peak at 0.45 FRET increases (Figure 4.2C-D), suggesting a measurable equilibrium between free and bound DNA rather than saturation or irreversible binding. In support of this, single time trajectories show interconversion between bound and unbound states, providing access to rates of binding and dissociation. As expected for reversible complex formation, the peak at a FRET value of 0.45 can be DNA competed away by adding unlabelled DNA (Figure 4.2F-H). This is true when unlabeled DNA is added after labeled DNA is already bound and also when unlabeled DNA is mixed together with labeled DNA and added to free protein.

cTALEs activate transcription

To test if cTALEs activate transcription, we designed a reporter assay. Four cTALE repeats were inserted into a host TALE (PthXo1) attached to a yeast GAL4 transcription activation domain (GAL4 TAD). Ten PthXo1 cognate binding sites are added upstream a reporter gene (EGFP). Yeast were transformed with plasmids containing the transcriptional activator, cTALE~Gal4, and the reporter. The amount of reporter transcripts produced is monitored by mRNA extraction, cDNA preparation, and qPCR. Compared to a transcription activator protein containing only the GAL4 TAD, the cTALE transcriptional activator increases production of reporter mRNA (Figure 4.2E). This shows that the cTALEs are capable of activating transcription from in a sequence-specific format.

cTALE arrays display mutlphasic DNA-binding kinetics.

In addition to the short smTIRF movies used to generate smFRET histograms, long movies were also collected to examine the transitions between the low- and high-FRET states. These long time trajectories show many transitions between low and high FRET (0.0 to 0.45) states in traces of single molecules (Figure 4.3A-B). A transition from low to high FRET (0 to 0.45) indicates the acceptor fluorophore on DNA moved close enough to the protein to be excited by the donor fluorophore and is likely a binding event. A transition from high to low FRET (0.45 to 0.0) indicates the acceptor fluorophore on DNA moved too far away from the protein to be excited by the donor fluorophore and is likely an unbinding event. Low-FRET states show low Cy5 co-localization in alternating laser experiments confirming that high-FRET states are DNA-bound states and low-FRET states are DNA-free states (Figure 4.S2). These long single molecule traces show both long- and short-lived low- and high-FRET states indicating kinetics are multi-phasic (Figure 4.3A-B). Transitions to high FRET (binding events) become more frequent as bulk DNA concentration increases (compare representative traces at 1 nM dsDNA to 15 nm dsDNA; Figure 4.3A and Figure 4.3B). Cumulative distributions generated with all low FRET dwell times at a given DNA concentration are best-fit by a double-exponential decay, indicating a minimum of two kinetic phases associated with binding events (Figure 4.3C). Cumulative distributions generated with all high FRET dwell times at a given DNA concentration are best-fit by a double-

exponential decay, indicating that there are a minimum of two kinetic phases for unbinding as well (Figure 4.3D).

The DNA concentration dependence of the two rates associated with transitions from low to high FRET (0.0 to 0.45; binding events) shows the faster phase is DNA concentration dependent (Figure 4.3E), indicating that this step involves an associative binding mechanism, whereas the slower phase is independent of DNA concentration indicating a unimolecular isomerization mechanism (Figure 4.3E). The rate constant for this slow phase is 0.59 s^{-1}. Taking the slope of the apparent rate constant for the fast phase as a function of DNA gives a bimolecular rate constant of $5.9x10^{8}$ $nM^{-1}s^{-1}$, close to the diffusion limit.

In contrast, neither of the two fitted rate constants for transitions from high to low FRET (0.45 to 0.0; unbinding events) depends on DNA concentration, suggesting that uninding involves two unique unimolecular processes (Figure 4.3F). Apparent rate constants are calculated as the average rate of the fast and slow phases (1.2 s^{-1} and 0.13 s^{-1} respectively).

To rule out kinetic contribution of TALEs threading axially onto the ends of short DNAs, binding kinetics were measured with capped double-helical DNA sites. Capped DNA was generated by forming 5'digoxygenin-A_5-Cy5-A_{15} duplexed with 5'-digoxygenin-T_{26} and adding three-fold molar excess α-Digoxygenin. Low and high FRET dwell time cumulative distributions generated from capped DNA-binding kinetics are bi-phasic, similar to uncapped DNA. To assess the affect of molecular weight changes on diffusion of capped and

uncapped DNA, Sednterp was used to estimate maximum diffusion coefficients. Due to the two antibodies bound on the ends of capped DNA, the estimated diffusion coefficient of the 320 kDa capped DNA ($4.7x10^{-7}$ cm^2s^{-1}) is much slower compared to the estimated diffusion coefficient of the 10 kDa uncapped DNA ($1.5 x10^{-6}$ cm^2s^{-1}). The apparent bimolecular rate constant is much smaller for the capped DNA, consistent with the slower diffusion of capped DNA compared to uncapped DNA (Figure 4.S1).

Longer cTALEs bind DNA with slower binding and unbinding kinetics

To examine how increasing the length of the cTALE array influences DNA binding, we generated Cy3-labelled constructs with 16 and 12 rather than 8 cTALE repeats, and measured binding to a longer Cy5-labelled DNA (A_{23}/T_{23}). Because did not observe FRET, co-localization was measured.

Increasing the number of cTALE repeats from 8 to 12 or 16 dramatically affects DNA binding kinetics. Long movies collected over a range of bulk DNA concentrations show short- and long-timescale on and off events similar to $N(NS)_8$. Single traces were analyzed using thresholding to identify states and dwell times. Cumulative distributions were generated from low red-channel emission (unbound states, with lifetimes representing binding events), and for high red-channel emission (bound states, with dwell times representing unbinding events). Whereas unbound cumulative distributions at low DNA concentration are best-fit by single exponential decays, those above 500 pM DNA are best-fit by double exponential decays. Bound cumulative distributions

are best-fit by double exponential decays. All apparent rate constants are much smaller for $N(NS)_{16}$ and $N(NS)_{12}$ (green/black circles and triangles, Figure 4.4A-B), indicating that binding and/or unbinding is impeded by increasing the length of the binding surface between cTALEs and their cognate DNA.

A deterministic approach to modeling cTALE-DNA binding kinetics.

To figure out how the kinetic changes above are partitioned into underlying kinetic steps in binding, we fitted various kinetic models to the cumulative distributions for binding and unbinding. In addition to providing information about the mechanism of binding, this approach allows us to estimate the underlying microscopic rate constants for binding and unbinding. Numerical integration was used to calculate the concentration of cTALE states as a function of time (Figures 4.5A-C and 4.5G-H), given a binding mechanism, an associated set of and rate laws, and a set of initial conditions. cumulative distributions of low FRET (0.0) dwell times represent the distribution of times single molecules spent in the low FRET state before transitioning into the high FRET (0.45) state. For this reason, it makes sense to split the kinetic scheme when fitting to single-molecule dwell times. Simulated events before a transition from low to high FRET (0.0 to 0.45; binding reactions) are considered together. Simulated events before a transition from high to low FRET (0.45 to 0.0; unbinding reactions) are considered together.

Among the various models tested, the model that is most consistent with the data has two low FRET DNA-free states, and two high FRET DNA-bound

states. This is consistent with alternating laser experiments showing DNA is only co-localized when cTALEs are in the high FRET state (Figure 4.S2). This model includes a TALE isomerization step in the absence of DNA, from a DNA-binding incompetent conformation to DNA-binding competent conformation (which we refer to as TALE*). The DNA-binding competent TALE* conformer binds and unbinds DNA (called TALE* when DNA-free and TALE*~DNA when DNA-bound). Before unbinding, a fraction of TALE*~DNA isomerizes to a longer-lived DNA-bound state called TALE~DNA.

Based on this mechanism, the rate laws for binding are given in equations 1a - 1d.

$$\frac{d[TALE]}{dt} = -k_1[TALE] + k_{-1}[TALE^*] \tag{1a}$$

$$\frac{d[TALE^*]}{dt} = k_1[TALE] - k_{-1}[TALE^*] - k_2[TALE^*][DNA] \tag{1b}$$

$$\frac{d[TALE^* \sim DNA]}{dt} = k_2[TALE^*][DNA] \tag{1c}$$

$$K_{eq,DNA-free} = \frac{k_1}{k_{-1}} \tag{1d}$$

To determine microscopic rate constants k_1, k_{-1}, and k_2, equations 1a-1c were numerically integrated in Matlab, and the fraction of TALE*~DNA as a function of time was fitted to the low-FRET cumulative distributions ($cTALE_8$; Figure 4.5D-E) or to the low co-localization cumulative distributions ($cTALE_{16}$; Figure 4.5J-K). Microscopic rate constants were adjusted to reduce sum of the squared residuals between the concentration of TALE*~DNA as a function of time and single-molecule cumulative distributions. In both cases, cumulative distributions at

different bulk DNA concentrations were fitted globally. Initial concentrations of TALE and TALE* conformers were determined by the fitted values of k_1 and k_{-1}, assuming a rapid pre-equilibrium; the initial fraction of TALE*~DNA was set to 0. Confidence intervals were set to 95% and estimated by 2000 bootstrap iterations (Table 4.1; best-fit and 95% CI).

Rate laws for dissociation are given in equations 2a - 2d

$$\frac{d[TALE^* \sim DNA]}{dt} = -k_{-2}[TALE^* \sim DNA] - k_3[TALE^* \sim DNA] + k_{-3}[TALE \sim DNA] \quad (2a)$$

$$\frac{d[TALE \sim DNA]}{dt} = k_3[TALE^* \sim DNA] - k_{-3}[TALE \sim DNA] \quad (2b)$$

$$\frac{d[TALE^*]}{dt} = k_{-2}[TALE^* \sim DNA] \quad (2c)$$

$$K_{eq,DNA-bound} = \frac{k_3}{k_{-3}} \quad (2d)$$

To determine microscopic rate constants k_{-2}, k_{-3}, and k_3, equations 2a-2c were numerically integrated in Matlab, and the fraction of TALE* as a function of time was fitted to the high-FRET cumulative distributions ($cTALE_8$; Figure 4.5F) or to the low co-localization cumulative distributions ($cTALE_{16}$; Figure 4.5L). Microscopic rate constants were adjusted to reduce sum of the squared residuals between the concentration of TALE* as a function of time and single-molecule cumulative distributions. In both cases, cumulative distributions at different bulk DNA concentrations were fitted globally. Initial fraction of TALE*~DNA conformer was set at 1; all other initial fraction were set to 0 (according to law of mass

action) Confidence intervals were set to 95% and estimated by 2000 bootstrap iterations (Table 4.1; best-fit and 95% CI).

Fitted curves are compared with measured cumulative distributions in Figure 4.4. Fitted curves reproduce the experimental cumulative distributions for binding and unbinding, both for the short and long cTALE arrays with reasonably small residuals, over a range of DNA concentrations. Generally, fitted rate constants have confidence intervals of 10% or smaller.

Comparison of micro rate constants for 8 and 16 repeats show some similarities but some important differences. Similar is the bimolecular microscopic binding rate constant, k_2, (1.76 and 1.3 $nM^{-1}s^{-1}$ for 8 and 16 repeats respectively), which makes sense due to the diffusion limit. But microscopic unbinding rate constant, k_{-2}, is larger for 8 repeat cTALEs (0.65 s^{-1} for $N(NS)_8$ versus 0.26 s^{-1} for $N(NS)_{16}$). Also, bound state isomerization (interconversion between TALE*~DNA and TALE~DNA) is 3-5 times slower for 16 repeat cTALEs than 8 repeat cTALEs. One other difference is initial fractions of the DNA-binding competent TALE*, which are greater than DNA-binding incompetent TALE in for cTALEs with 8 repeats ($K_{eq,\ DNA\text{-}free} = 6.0$), while initial concentration of DNA-binding competent TALE* is less than DNA-binding incompetent TALE in cTALEs with 16 repeats ($K_{eq,\ DNA\text{-}free} = 0.6$).

4.4 Discussion

cTALEs containing NS RVD bind DNA with high affinity

NS is an uncommon RVD in natural TALEs. Previous reports suggest it is fairly nonspecific, but may bind with higher affinity than other common RVDs (NG, NI, NN, and HD)[5]. Our results show that cTALEs containing the NS RVD bind DNA very tightly with an apparent K_d (K_{app}) calculated from using equation 3.

$$K_{app} = \frac{[TALE - DNA + TALE^* - DNA]}{[DNA][TALE + TALE^*]} \tag{3a}$$

$$K_1 = \frac{k_1}{k_{-1}} \tag{3b}$$

$$K_2 = \frac{k_2}{k_{-2}} \tag{3c}$$

$$K_3 = \frac{k_3}{k_{-3}} \tag{3d}$$

$$K_{app} = \frac{K_1K_2 + K_1K_2K_3}{1 + K_1} \tag{3e}$$

$$K_{app} = \frac{\frac{k_1}{k_{-1}} \times \frac{k_2}{k_{-2}}(1 + \frac{k_3}{k_{-3}})}{1 + \frac{k_1}{k_{-1}}} \tag{3f}$$

K_{app} is 5.67 nM for the 8 repeat cTALE array and 4.13 nM for the 16 repeat cTALE array. Doubling the number of repeats has a modest affect on the apparent K_d due to the increased population of binding incompetent DNA-free TALE in the 16 repeat array. This affinity change is much less than a previous report studying length dependence on affinity of designed TALEs (dTALEs) showing the K_d of a dTALE decreased by a factor of two with the addition of only 1.5 repeats[21].

High affinity binding of cTALE arrays containing the NS RVD is specific to adenine bases. Due to the labeling position on the cTALE and the length of shorter DNA used to study binding of the 8 repeat cTALEs, models of distance between donor and acceptor dyes are similar if the 8 repeat cTALE binds in the adenine-sense or thymine-sense orientation. However, because there is co-localization but no observed FRET when 16 repeat NS RVD cTALEs bind DNA labeled with Cy5 on the 5'-end of the A_{23} strand, the binding orientation and therefore nucleotide preference can be determined. Consistent with previous reports, cTALEs containing the NS RVD prefer adenine compared with thymine bases.

Because the N-terminal capping repeats in our constructs have been shown to make little contribution to DNA binding affinity, of the binding energy in our constructs is likely to derive from the cTALE repeats. Also, FRET between $N(NS)_8$ and Cy5-labeled G_{15}/C_{15} DNA was not observed at 200 mM KCl. FRET is observed when 16 repeat cTALEs containing the HD RVD ($N(HD)_{16}$) are incubated with 50 nM Cy5-labeled G_{20}/C_{20} at 50 mM KCl. DNA binding of 16 repeat cTALEs containing the HD RVD was too weak to perform dwell time analysis due to background from high concentrations of DNA required to detect.

Conformational heterogeneity in the unbound state could be local unfolding

Although the deterministic modeling is useful for testing different kinetic models and determining microscopic rate constants as well as equilibrium

constants, such analysis does not provide information about the structural nature of TALE conformational heterogeneity. There is clear conformational heterogeneity in both free TALEs and DNA-bound TALEs (Figure 4.5). For 8 repeat cTALEs, the DNA-binding competent state is more highly populated than the DNA-binding incompetent state. In this reaction scheme, the DNA-binding incompetent state can be regarded as an off-pathway conformation that inhibits DNA binding (Figure 4.6A). For 16 repeat cTALEs, the DNA-binding incompetent state is more highly populated than the DNA-binding competent state. In this reaction scheme, the DNA-binding competent state is an on-pathway intermediate required for DNA binding (Figure 4.6B).

The population of end frayed states is similar to the population of the DNA-binding incompetent state of 8 repeat cTALEs. Although the population of partly folded states is predicted to be less than the population of DNA-binding competent state of 16 repeat cTALEs, it is also possible, that many partly folded or broken conformations are included in what we call the DNA-binding competent state. Figure 4.1 shows a single broken conformation for simplicity, but breaks are equally likely to occur at each interface between repeats. The chance of that break happening at any position in a repeat array increases linearly with the number of repeats in the array.

Assuming the DNA-binding incompetent state of 8 repeat cTALEs is partly unfolded, microscopic rate constant k_1 corresponds to local folding (0.13 s^{-1}). This rate constant is similar the folding rate constant of 16 repeat cTALEs, microscopic rate constant k_{-1} (0.1 s^{-1}). The microscopic unfolding rate constant

indicates a decrease in unfolding rate of 8 repeat cTALEs (0.022 s^{-1}) compared to 16 repeat cTALEs (0.06 s^{-1}). Because there are more repeats available to unfold in a longer array, the apparent rate constant of unfolding likely increases.

Short TALEs don't locally unfold to bind DNA

Figure 4.6 shows a model of conformational change consistent with DNA binding kinetics. In this model there are four TALE states. Closed and open represent the two DNA-free states. Encounter and locked represent the two DNA-bound states. Because the cTALE8 array does not form multiple turns of a superhelix, unfolding to bind DNA is not required. The closed state is active to bind DNA. However, because 16 repeat cTALEs form 1.4 turns (excluding the N-terminal domain) unfolding to bind DNA is required. The open state is active to bind DNA.

Increasing populations of partly folded states through addition of 1M urea and entropy enhancing mutations decreases apparent binding rates of 8 repeat cTALEs (Figure 4.S3). This is also consistent with a partly folded DNA-binding incompetent state in shorter cTALE arrays.

Conformational heterogeneity in the bound state

Previous reports show that TALEs have multiple diffusional modes when searching nonspecific DNA[22]. Our work suggests that cTALEs have multiple binding modes (encounter and locked states in Figure 4.6). cTALEs undergo a conformational change also when bound to specific DNA sequences. Table 4.1

shows that microscopic rate constants for transition into and out of longer lived locked bound states become much slower in 16 repeat cTALEs compared with 8 repeat cTALEs (k_{-3} and k_3). These rate constants decrease much more than the microscopic unbinding rate constant (the k_{-2} values are 0.65 s^{-1} and 0.26 s^{-1} for $N(NS)_8$ and $N(NS)_{16}$ respectively) suggesting a large conformational change dependent on the number of repeats. Although the model does not provide information on structure of this conformational change, it is possible this conformational change involves a slinky motion to decrease helical rise (still 11.5 repeats per turn, but rise decreases upon binding) or specific interaction with RVDs and bases in the major groove of DNA. Crystal structures show little deformation of DNA structure, so bending of DNA seems unlikely. cTALEs bind DNA through short encounters which occasionally become long-lived locked conformations (Figure 4.6).

4.5 Materials and Methods

Cloning, expression, purification, and labeling

Consensus TALE repeat constructs were cloned with C-terminal His_6 tags via an in-house version of Golden Gate cloning[23]. TALE constructs were grown in BL21(T1R) cells at 37°C to an OD of 0.6-0.8 and induced with 1 mM IPTG. Following cell pelleting and lysis, proteins were purified by resuspending the insoluble material in 6M urea, 300 mM NaCl, 0.5 mM TCEP, and 10 mM $NaPO_4$ pH 7.4. Constructs were loaded onto an Ni-NTA column. Protein was eluted

using 250 mM imidazole and refolded during buffer exchange into 300 mM NaCl, 30% glycerol, 0.5 mM TCEP, and 10 mM $NaPO_4$ pH 7.4.

Labelling of cTALE arrays followed a previously reported protocol[24]. 1 mg protein is loaded onto 500 uL NiNTA spin column. The column is washed with 10 column volumes of 300 mM NaCl, 0.5 mM TCEP, and 10 mM $NaPO_4$ pH 7.4 buffer. Tenfold molar excess Cy3 maleimide dye is resuspended in 10 μL DMSO and added to column. The column rocks at room temperature for 30 minutes, then at 4°C overnight. Cy3-labeled protein eluted with 250 mM imidazole, 300 mM NaCl, 30% glycerol, 0.5 mM TCEP, and 10 mM $NaPO_4$ pH 7.4. Protein was stored in 300 mM NaCl, 30% glycerol, 0.5 mM TCEP, and 10 mM $NaPO_4$ pH 7.4 buffer at -80°C.

Oligonucleotides

Sequences used for binding studies were 5'-Cy5-A_{15}-3' and 5' T_{15}-3' duplex (Cy5-A_{15}/T_{15}) for 8 repeat binding studies, and 5'-Cy5-A_{23}-3' and 5' T_{23}-3' duplex (Cy5-A_{23}/T_{23}) for 16 repeat binding studies. DNA was annealed at 5 μM concentration with 1.2-fold molar excess unlabeled strand in 10 mM Tris pH 7.0, 30 mM NaCl.

Single-molecule detection and data analysis

Biotinylated quarts slides and glass coverslips were prepared as previously described[24]. Cy3-labeled cTALEs were immobilized on biotinylated slides taking advantage of neutravidin interaction with biotinylated α-penta•His which binds the His_6 cTALE tag. Slides are pretreated with blocking buffer (5 μL yeast tRNA, 5 μL BSA, 40 μL T50) before addition of 250 pM labeled cTALE.

Cy5-labeled DNA is mixed with imaging buffer (20 mM Tris pH 8.0, 200 mM KCl, 0.5 mg mL^{-1} BSA, 1 mg mL^{-1} glucose oxidase, 0.004 mg mL^{-1} catalase, 0.8% dextrose and saturated Trolox ~1mg mL^{-1}) and molecules were imagined using total internal reflection fluorescence microscopy. Time resolution was 50 msec for $N(NS)_8$ and 100 msec for $N(NS)_{816}$. Collection and analysis performed as previously described[25].

FRET histograms

A minimum of 20 short movies were collected, and the first 5 frames (50 msec exposure time) were used to generate smFRET histograms. FRET calculated as $I_A/(I_A+I_D)$ where IA and ID are donor-leakage and background corrected fluorescence emission of acceptor (Cy5) and donor (Cy3) fluorophores. In competition experiments, unlabeled DNA with the same sequence as labeled DNA was mixed at indicated concentrations with labeled DNA prior to imaging.

***S. cerevisiae* transcription activation assay**

Reporter assay includes two separate plasmids: (1) DNA-binding domains fused to GAL4-TA as one transcript with self-cleaving P2 peptide and dsRED (backbone vector pAG415GPD-ccdB-DsRed) and (2) reporter plasmid containing 10 TALE cognate binding sites (5'- TGCATCTCCCCCTACTGTACACCAC -3') imbedded in the a synthetic minimal promoter[26] controlling expression of EGFP (backbone vector pAG416GAL-EGFP-ccdB). DNA-binding domains studied include the naturally occurring TALE, PthXo1 including a mutation, A32D, in the fourth repeat to accommodate cloning (WT), PthXo1 constructs where first four repeats are replaced with cTALE repeats containing the NS RVD (NS_4), PthXo1

constructs where first four repeats are replaced with cTALE repeats containing the HD RVD (HD_4), and the control with no DNA-binding domain (control) (Figure 4.S4).

Yeast are transformed with plasmids (1) and (2) and grown under minimal selection at 30°C. Cultures grown to saturation overnight, diluted to a starting OD of 0.2 and grown four more hours before pelleting cells and storing at -80°C. Cells lysed and mRNA acid phenol-chloroform extracted followed by DNase treatment, and subsequent cDNA preparation. Reporter mRNA levels quantified by qRT-PCR using iQ5 iCycler system (Bio-Rad, Hercules, CA) and iQ SYBR Green Supermix (Bio-Rad, Hercules, CA) using qRT primers designed against reporter construct EGFP and transformation control dsRED.

Dwell time analysis

Long movies collected with 50 msec exposure time for $cTALE_8$ and 100 msec exposure time for $cTALE_{16}$. At least 20 representative traces at each DNA concentration are selected and dwell times are determined by fitting as previously described using HaMMy[27] for FRET in $cTALE_8$. Dwell times are determined by thresholding of Cy5 excitation for $cTALE_{16}$ co-localization experiments. All FRET and co-localization data are well described by models with 2 distinct states (0.0 FRET and ~0.45 FRET as well as low co-localization and high co-localization. Dwell times of the same state (low versus high FRET or low versus high co-localization) for all traces at a given DNA concentration are compiled, and cumulative distribution is generated with spacing equal to imaging exposure time.

To determine apparent rate constants using model-independent analysis, cumulative distributions are fitted with single and double exponential decays (Figure 4.3 and 4). Observed rates from exponential decay fits are plotted as a function of DNA concentration. Apparent rate constants calculated as slope of DNA concentration-dependent observed rates or average of DNA concentration-independent observed rates.

Deterministic modeling

Equations 1a-1c and 2a-2c were numerically integrated using ODE15s and ODE45 solver in MATLAB. Microscopic rate constants were adjusted to minimize the sum of squared residuals between ODE-determined concentration of bound or free TALE and single molecule cumulative distributions using lsqnonlin in MATLAB. 95% confidence intervals were estimated by performing 2000 bootstrap iterations.

4.6 References

1. Kay, S., Hahn, S., Marois, E., Hause, G. & Bonas, U. A bacterial effector acts as a plant transcription factor and induces a cell size regulator. *Science* **318,** 648–651 (2007).
2. Römer, P. *et al.* Plant pathogen recognition mediated by promoter activation of the pepper Bs3 resistance gene. *Science* **318,** 645–648 (2007).
3. Boch, J. *et al.* Breaking the code of DNA binding specificity of TAL-type III effectors. *Science* **326,** 1509–1512 (2009).
4. Moscou, M. J. & Bogdanove, A. J. A simple cipher governs DNA recognition by TAL effectors. *Science* **326,** 1501 (2009).
5. Miller, J. C. *et al.* Improved specificity of TALE-based genome editing using an expanded RVD repertoire. *Nat. Methods* **12,** 465–471 (2015).
6. Cong, L., Zhou, R., Kuo, Y.-C., Cunniff, M. & Zhang, F. Comprehensive interrogation of natural TALE DNA-binding modules and transcriptional repressor domains. *Nat. Commun.* **3,** 968 (2012).
7. Geissler, R. *et al.* Transcriptional activators of human genes with programmable DNA-specificity. *PloS One* **6,** e19509 (2011).
8. Li, Y., Moore, R., Guinn, M. & Bleris, L. Transcription activator-like effector hybrids for conditional control and rewiring of chromosomal transgene expression. *Sci. Rep.* **2,** 897 (2012).
9. Mahfouz, M. M. *et al.* Targeted transcriptional repression using a chimeric TALE-SRDX repressor protein. *Plant Mol. Biol.* **78,** 311–321 (2012).

10. Zhang, F. *et al.* Efficient construction of sequence-specific TAL effectors for modulating mammalian transcription. *Nat. Biotechnol.* **29,** 149–153 (2011).

11. Maeder, M. L. *et al.* Targeted DNA demethylation and activation of endogenous genes using programmable TALE-TET1 fusion proteins. *Nat. Biotechnol.* **31,** 1137–1142 (2013).

12. Miyanari, Y., Ziegler-Birling, C. & Torres-Padilla, M.-E. Live visualization of chromatin dynamics with fluorescent TALEs. *Nat. Struct. Mol. Biol.* **20,** 1321–1324 (2013).

13. Ma, H., Reyes-Gutierrez, P. & Pederson, T. Visualization of repetitive DNA sequences in human chromosomes with transcription activator-like effectors. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 21048–21053 (2013).

14. Li, T. *et al.* TAL nucleases (TALNs): hybrid proteins composed of TAL effectors and FokI DNA-cleavage domain. *Nucleic Acids Res.* **39,** 359–372 (2011).

15. Christian, M. *et al.* Targeting DNA Double-Strand Breaks with TAL Effector Nucleases. *Genetics* **186,** 757–761 (2010).

16. Deng, D. *et al.* Structural basis for sequence-specific recognition of DNA by TAL effectors. *Science* **335,** 720–723 (2012).

17. Mak, A. N.-S., Bradley, P., Cernadas, R. A., Bogdanove, A. J. & Stoddard, B. L. The crystal structure of TAL effector PthXo1 bound to its DNA target. *Science* **335,** 716–719 (2012).

18. Boch, J. & Bonas, U. Xanthomonas AvrBs3 family-type III effectors: discovery and function. *Annu. Rev. Phytopathol.* **48,** 419–436 (2010).

19. Geiger-Schuller, K. & Barrick, D. Broken TALEs: Transcription Activator-like Effectors Populate Partly Folded States. *Biophys. J.* **111,** 2395–2403 (2016).

20. Gao, H., Wu, X., Chai, J. & Han, Z. Crystal structure of a TALE protein reveals an extended N-terminal DNA binding region. *Cell Res.* **22,** 1716–1720 (2012).

21. Rinaldi, F. C., Doyle, L. A., Stoddard, B. L. & Bogdanove, A. J. The effect of increasing numbers of repeats on TAL effector DNA binding specificity. *Nucleic Acids Res.* (2017). doi:10.1093/nar/gkx342

22. Cuculis, L., Abil, Z., Zhao, H. & Schroeder, C. M. Direct observation of TALE protein dynamics reveals a two-state search mechanism. *Nat. Commun.* **6,** 7277 (2015).

23. Cermak, T. *et al.* Efficient design and assembly of custom TALEN and other TAL effector-based constructs for DNA targeting. *Nucleic Acids Res.* **39,** e82 (2011).

24. Rasnik, I., Myong, S., Cheng, W., Lohman, T. M. & Ha, T. DNA-binding Orientation and Domain Conformation of the E.coli Rep Helicase Monomer Bound to a Partial Duplex Junction: Single-molecule Studies of Fluorescently Labeled Enzymes. *J. Mol. Biol.* **336,** 395–408 (2004).

25. Roy, R., Hohng, S. & Ha, T. A practical guide to single-molecule FRET. *Nat. Methods* **5,** 507–516 (2008).

26. McKinney, S. A., Joo, C. & Ha, T. Analysis of single-molecule FRET trajectories using hidden Markov modeling. *Biophys. J.* **91,** 1941–1951 (2006).

27. Redden, H. & Alper, H. S. The development and characterization of synthetic minimal yeast promoters. *Nat. Commun.* **6,** 7810 (2015).

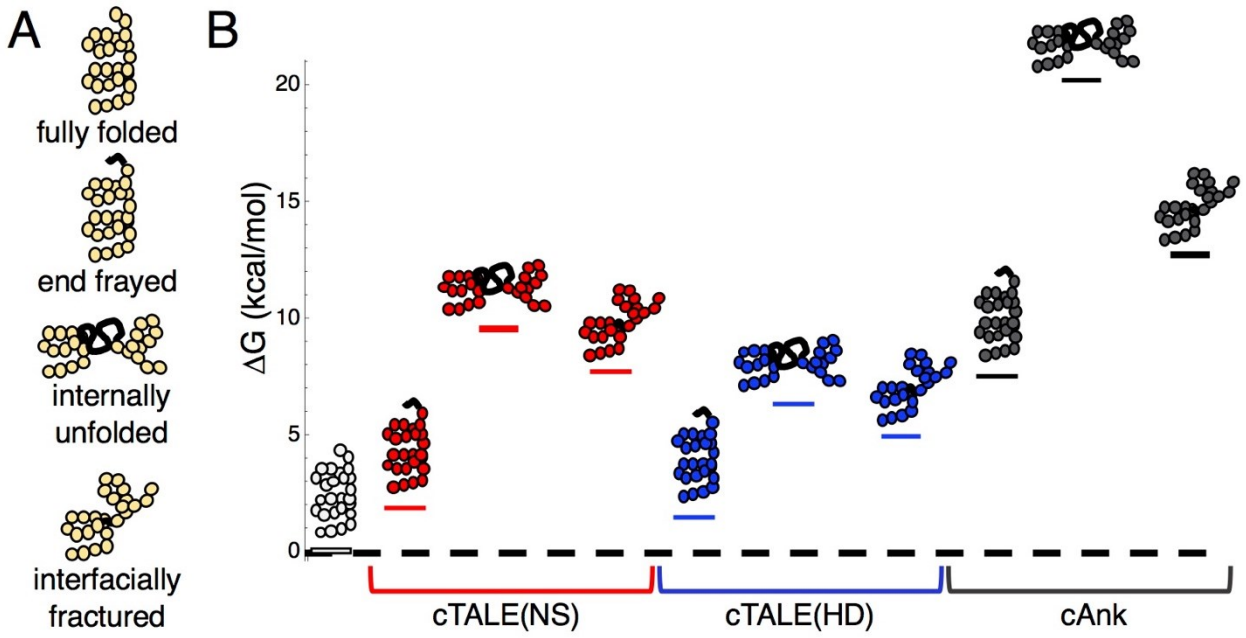


Figure 4.1. cTALEs populate partly folded states. (A) Cartoon of different partly folded TALE conformational states. End-frayed states have a terminal repeat unfolded. Internally unfolded states have a central repeat unfolded. "Fractured" states have a disrupted interface between adjacent repeats. (B) Calculated free energy of partly folded states for consensus TALE repeats with the NS repeat-variable diresidue sequence (cTALE(NS), red), consensus TALE repeats with the HD repeat-variable diresidue sequence (cTALE(HD), blue), and consensus ankyrin repeats(cAnk, black).

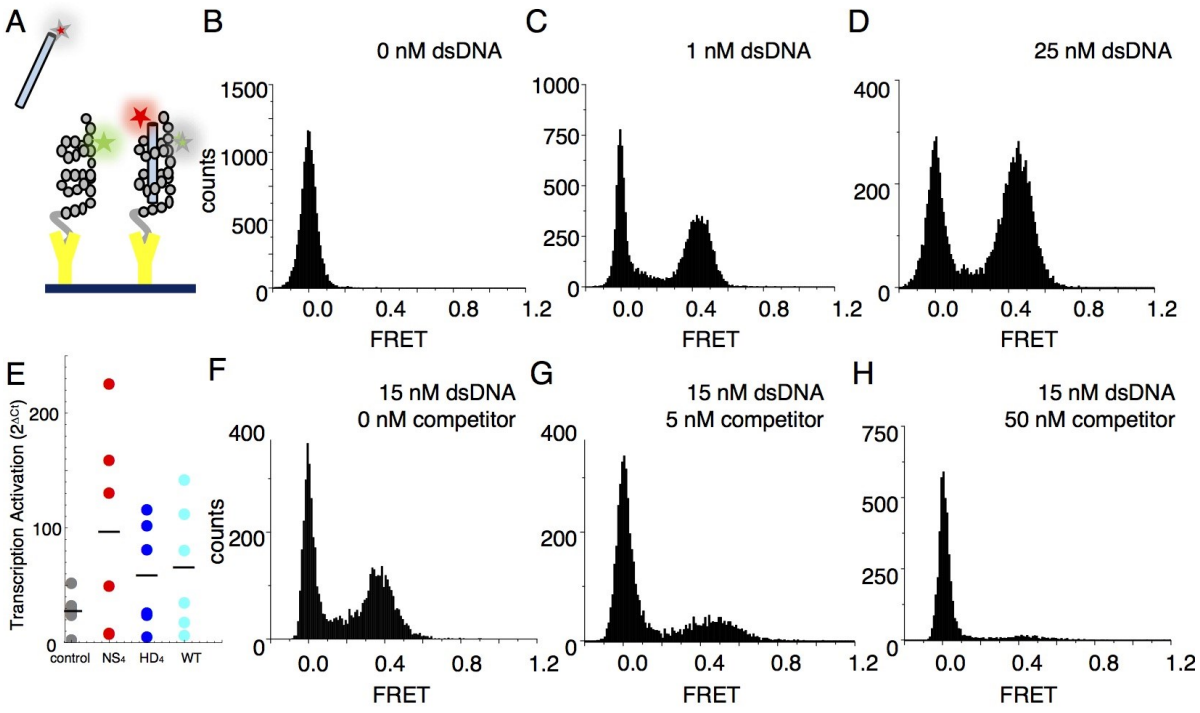


Figure 4.2. cTALEs bind dsDNA and activate transcription. (A) Schematic of single-molecule FRET assay, with donor-labelled cTALE attached to a surface, and acceptor-labelled DNA free in solution. (B-D) Single molecule FRET histograms show appearance of 0.45 FRET peak with increasing DNA. (E) Transcription activation assays show cTALEs activate transcription in *S. cerevisiae* cells. Compared to the control, cTALEs (NS_4 and HD_4) and a wild type TALE, PthXo1 (WT), increase the amount of reporter mRNA in qPCR experiments. Each point represents the average activation of a biological replicate measured in triplicate by qPCR. A black line shows the mean activation of all biological replicates. (F-H) SM FRET histograms show the addition of unlabeled DNA displaces labeled DNA. Conditions: 20 mM Tris pH 8.0, 200 mM KCl.

Figure 4.3. Single Molecule (SM) kinetics show multiple phases in binding and unbinding kinetics. (A-B) Time trajectories showing transitions between low- and high-FRET states (efficiencies of 0 and 0.45). The top panel shows calculated FRET efficiency in blue and two-state Hidden Markov Model fit in green[27]. The bottom panels show Cy3 and Cy5 fluorescence emission shown below in green and red respectively. At low DNA concentration (A), the low FRET state predominates. As DNA concentration is increased (B), more time is spent in the high FRET state, because the dwell times in the low FRET state are shorter. At low DNA concentrations, there appears to be long- and short-lived high-FRET states. Likewise, at near-saturating DNA concentrations, there appear to be long and short-lived low FRET states. (C, D) CDFs of low- and high-FRET dwell times. Fits to single-exponentials (black) show nonrandom residuals (lower panels), consistent with the heterogeneity noted in (A) and (B). Double-exponentials (red) give more uniform residuals. (E) Apparent association rate constants as a function of DNA concentration. The apparent rate constants for the fast phase are DNA concentration dependent (blue circles), indicating a bimolecular step binding event. The apparent rate constants for the slow phase do not depend on DNA-concentration (blue triangles), suggesting an isomerization event. (F) Apparent dissociation rate constants as a function of DNA concentration (phase 1 shown in blue circles, and phase 2 shown in blue triangles). Neither phase shows a DNA concentration dependence, indicating a dissociation and/or isomerization events. Conditions: 20 mM Tris pH 8.0, 200 mM KCl.

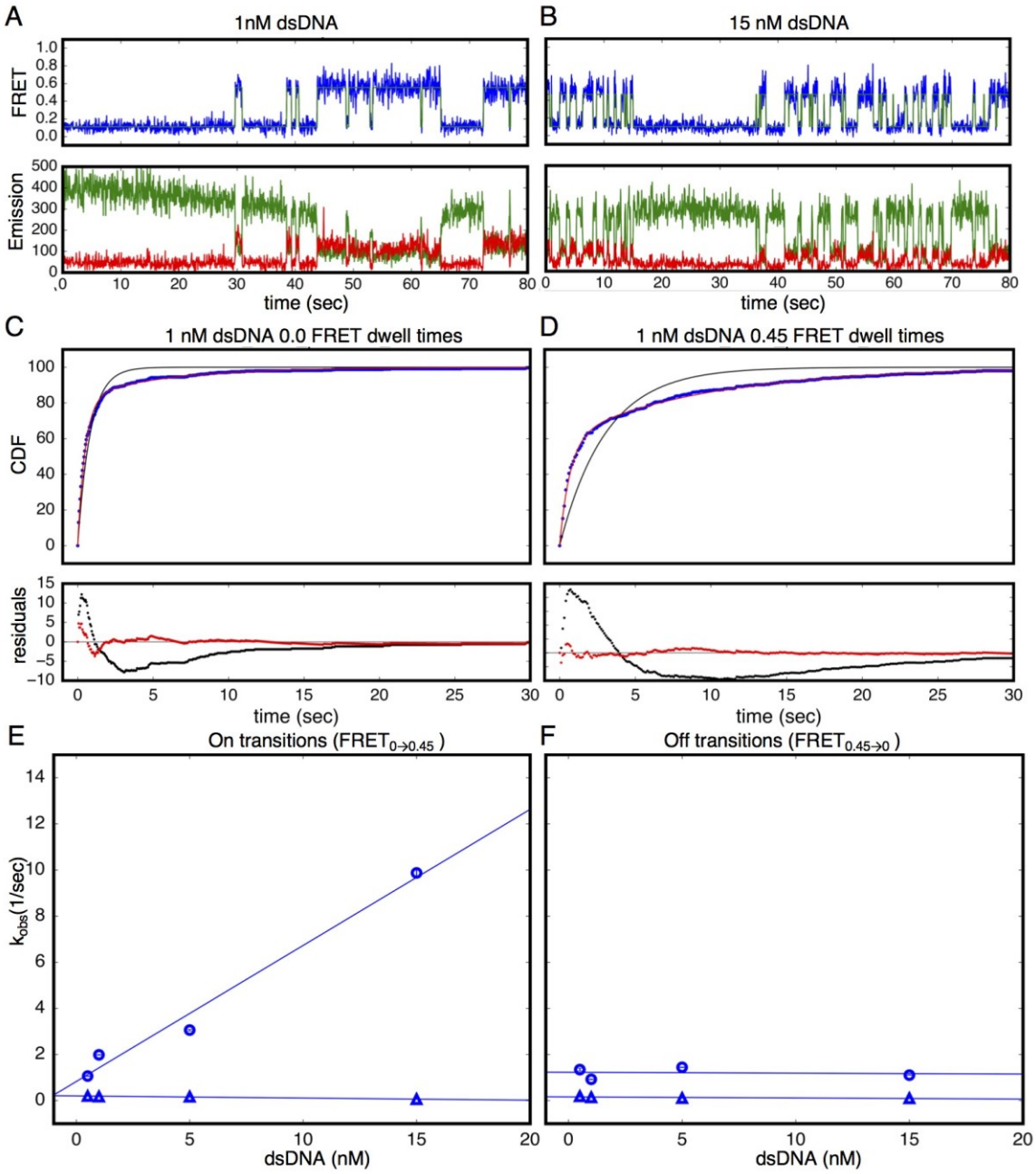

A
1nM dsDNA
B
15 nM dsDNA
FRET
Emission
time (sec)
C
1 nM dsDNA 0.0 FRET dwell times
D
1 nM dsDNA 0.45 FRET dwell times
CDF
residuals
time (sec)
E
On transitions (FRET0→0.45)
F
Off transitions (FRET0.45→0)
kobs(1/sec)
dsDNA (nM)

Figure 4.4. A 16-repeat TALE protein binds and unbinds DNA more slowly than an eight repeat protein. (A) Apparent association rate constants as a function of DNA concentration for an 8 repeat cTALE (blue), a 12 repeat cTALE (black), and a 16 repeat cTALE (green). 8 repeat TALE kinetics are measured by FRET ($FRET_{L\rightarrow H}$) while 12 and 16 repeat TALE kinetics are measured by co-localization ($E_{L\rightarrow H}$). The apparent rate constants for the fast phase are DNA concentration dependent (blue, black, and green circles), indicating a bimolecular step binding event. The DNA concentration-dependence is strongest (larger slope) for the 8 repeat cTALE. The apparent rate constants for the slow phase do not depend on DNA-concentration (blue, black, and green triangles), suggesting an isomerization event. (B) Apparent dissociation rate constants as a function of DNA concentration (phase 1 shown in circles, and phase 2 shown in triangles). Neither phase shows a DNA concentration dependence, indicating a dissociation and/or isomerization events. Rate constants for all phases are slower for the 12-repeat construct (black) and 16-repeat construct (green) than for the 8-repeat construct (blue), particularly for the bimolecular binding step. (C) Log_{10} of rate constants for 8 (blue), 12 (black), and 16(green) repeat cTALEs. Units of the bimolecular binding rate constant are $nM^{-1}s^{-1}$, while all other unimolecular rate constants have units s^{-1}. Conditions: 20 mM Tris pH 8.0, 200 mM KCl.

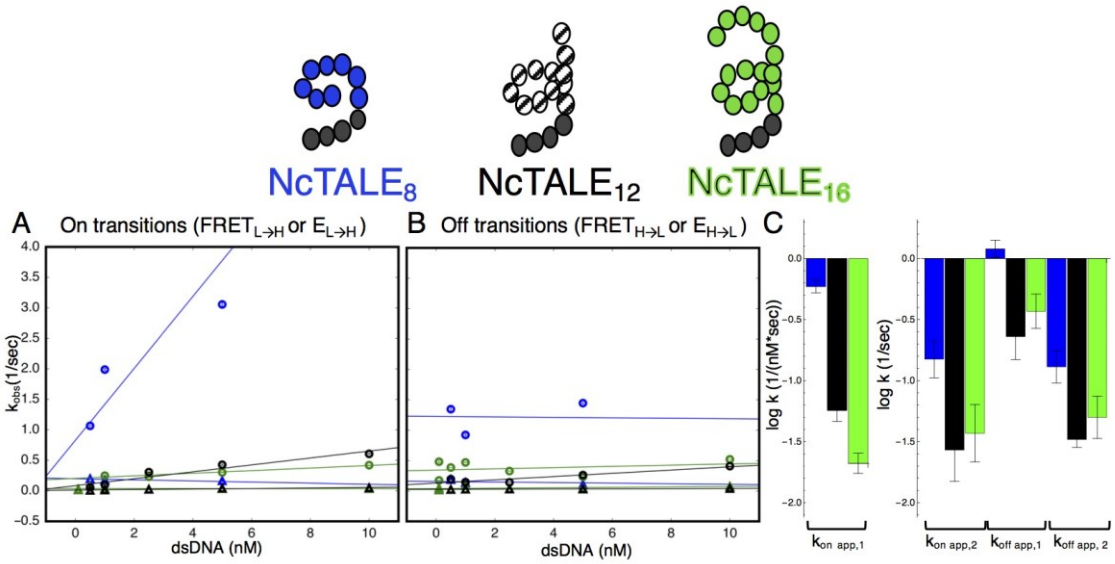
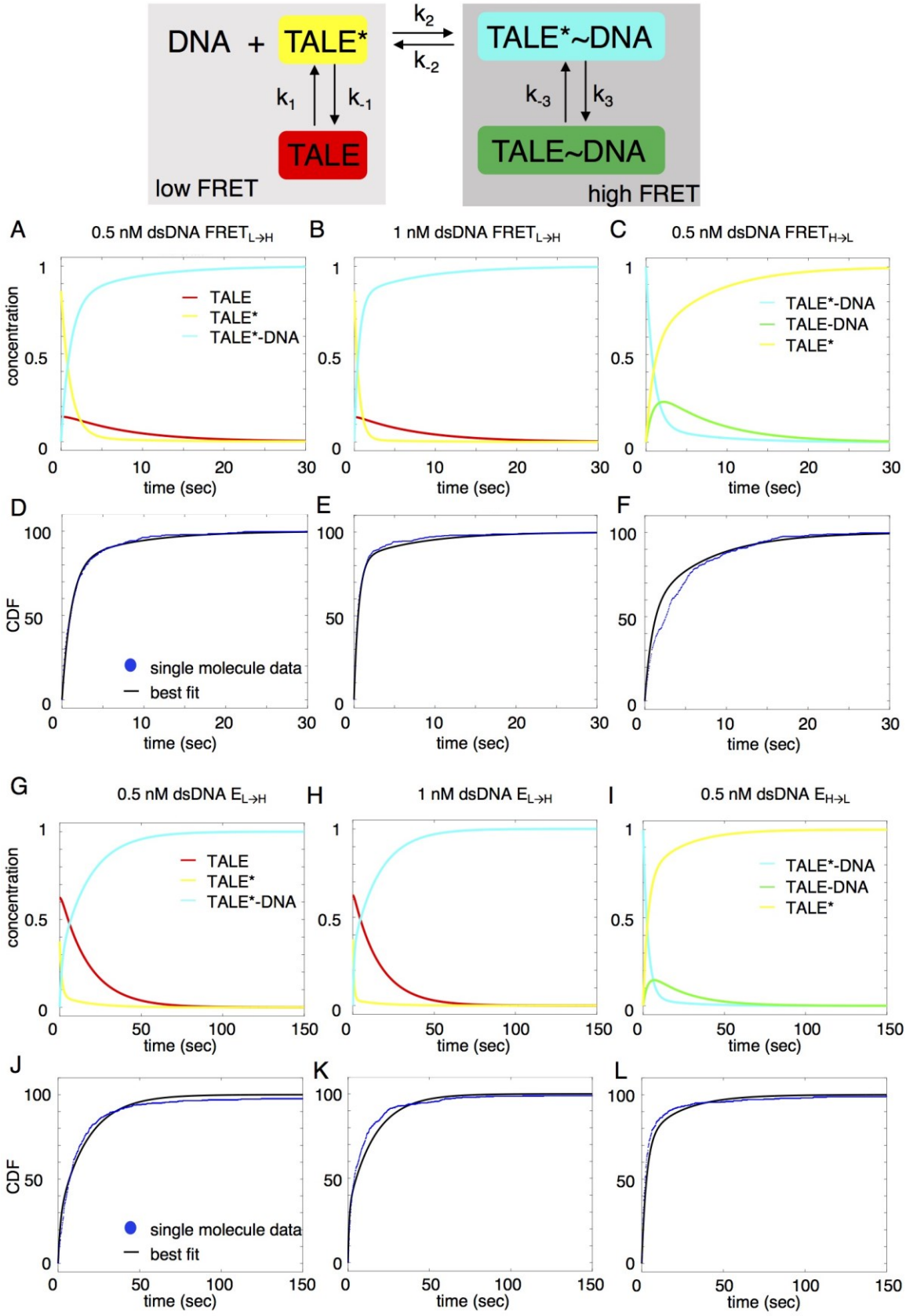

NcTALE8
NcTALE12
NcTALE16
A On transitions (FRET L→H or E L→H)
B Off transitions (FRET H→L or E H→L)
C
kobs (1/sec)
dsDNA (nM)
dsDNA (nM)
log k (1/(nM*sec))
log k (1/sec)
kon app,1
kon app,2
koff app,1
koff app, 2

Figure 4.5. Deterministic simulations provide evidence for conformational heterogeneity in the unbound state. The model most consistent with data is shown at the top. Unbound TALEs can exist in DNA-binding competent (TALE*) or DNA-binding incompetent (TALE) states. DNA-bound TALEs can exist in short-lived (TALE*~DNA) or long-lived (TALE~DNA) DNA-bound states. CDFs (shown as blue points) from 8 repeat single-molecule time trajectories (A-F) and 16 repeat single-molecule time trajectories (G-L) were analyzed with the model (best-fit shown in black). (A-C and G-I) Populations of states as a function of time, generated by numerical integration in Matlab. (D-F and J-L) CDF in blue circles and best fit lines shown in black. Best-fit microscopic rate constants and 95% confidence intervals shown in Table 4.1. For the 8-repeat cTALE construct, the DNA-binding competent (TALE*) state is more populated at t = 0. However, for 16 repeat cTALEs, the DNA-binding incompetent (TALE) state is more populated at t = 0.


DNA + TALE*
k1
k-1
TALE
low FRET
k2
k-2
TALE*~DNA
k-3
k3
TALE~DNA
high FRET
A
0.5 nM dsDNA FRET L→H
B
1 nM dsDNA FRET L→H
C
0.5 nM dsDNA FRET H→L
concentration
TALE
TALE*
TALE*-DNA
TALE-DNA
time (sec)
D
E
F
CDF
single molecule data
best fit
G
0.5 nM dsDNA E L→H
H
1 nM dsDNA E L→H
I
0.5 nM dsDNA E H→L
J
K
L

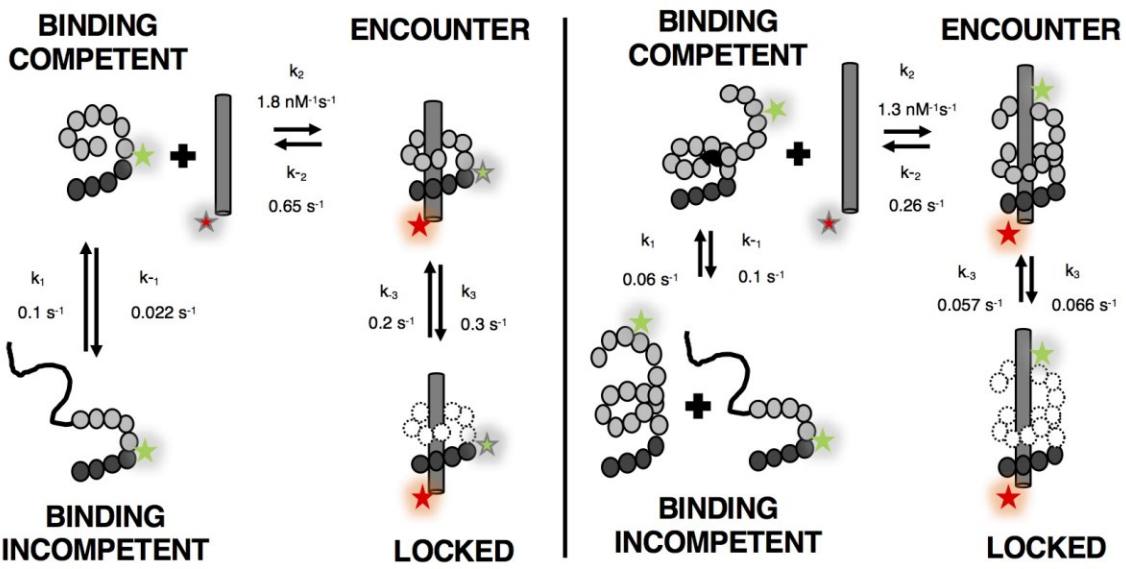


Figure 4.6. TALEs with multiple superhelical turns must break to bind DNA. Single-molecule FRET studies and deterministic modeling support a model where TALEs exist in four states: open, closed, encounter complex, and locked. (A) For shorter TALEs (8 repeats) that don't form multiple superhelical helical turns, partly folded states are off-pathway intermediates that slow down binding. (B) For longer TALEs (16 repeats) that form multiple superhelical turns, partly folded states are on-pathway intermediates required for binding. DNA-bound TALEs exist in encounter complexes or higher-affinity locked conformations. The time it takes for TALEs to enter the locked state as well as transition back to the encounter complex is 3-5 times for TALEs with 16 repeats (B) compared to TALEs with 8 repeats (B).

Table 4.1. Kinetic parameters obtained from deterministic simulation fits.

	k_1 (sec^{-1})	k_{-1} (sec^{-1})	$K_{eq,\ DNA\text{-}free}$	k_2($sec^{-1}nM^{-1}$)	k_{-2} (sec^{-1})	k_3 (sec^{-1})	k_{-3} (sec^{-1})	$K_{eq,\ DNA\text{-}bound}$
$N(NS)_8$[a]	0.13 [0.128, 0.137]	0.0222 [0.021, 0.024]	5.97 [5.76, 6.14]	1.76 [1.74, 1.79]	0.65 [0.63, 0.66]	0.32 [0.30, 0.35]	0.22 [0.21, 0.23]	1.48 [1.42, 1.55]
$N(NS)_{16}$[a]	0.063 [0.0628, 0.0638]	0.1 [0.101, 0.107]	0.61 [0.59, 0.62]	1.29 [1.25, 1.34]	0.26 [0.258, 0.265]	0.066 [0.063, 0.068]	0.057 [0.055, 0.059]	1.14 [1.11, 1.16]

[a]95% confidence intervals shown in brackets are from 2,000 iterations of bootstrap analysis.

4.7 Supplemental Material

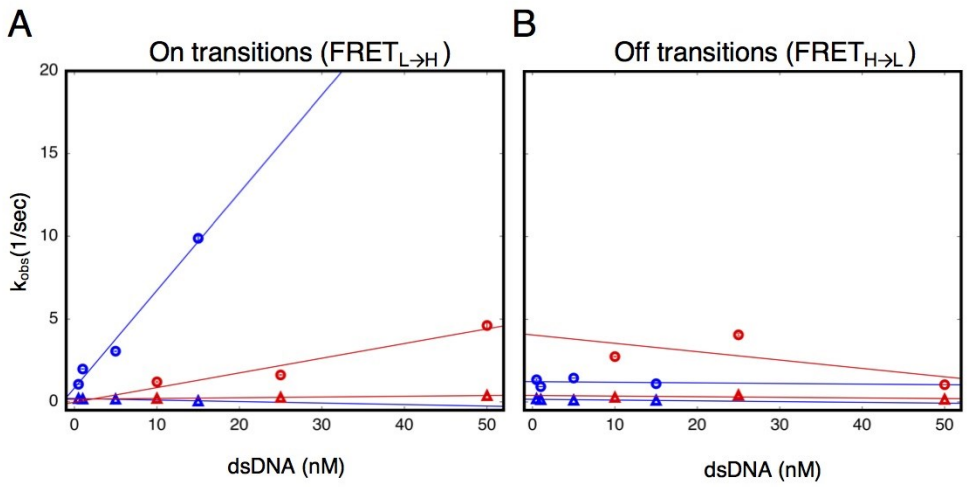
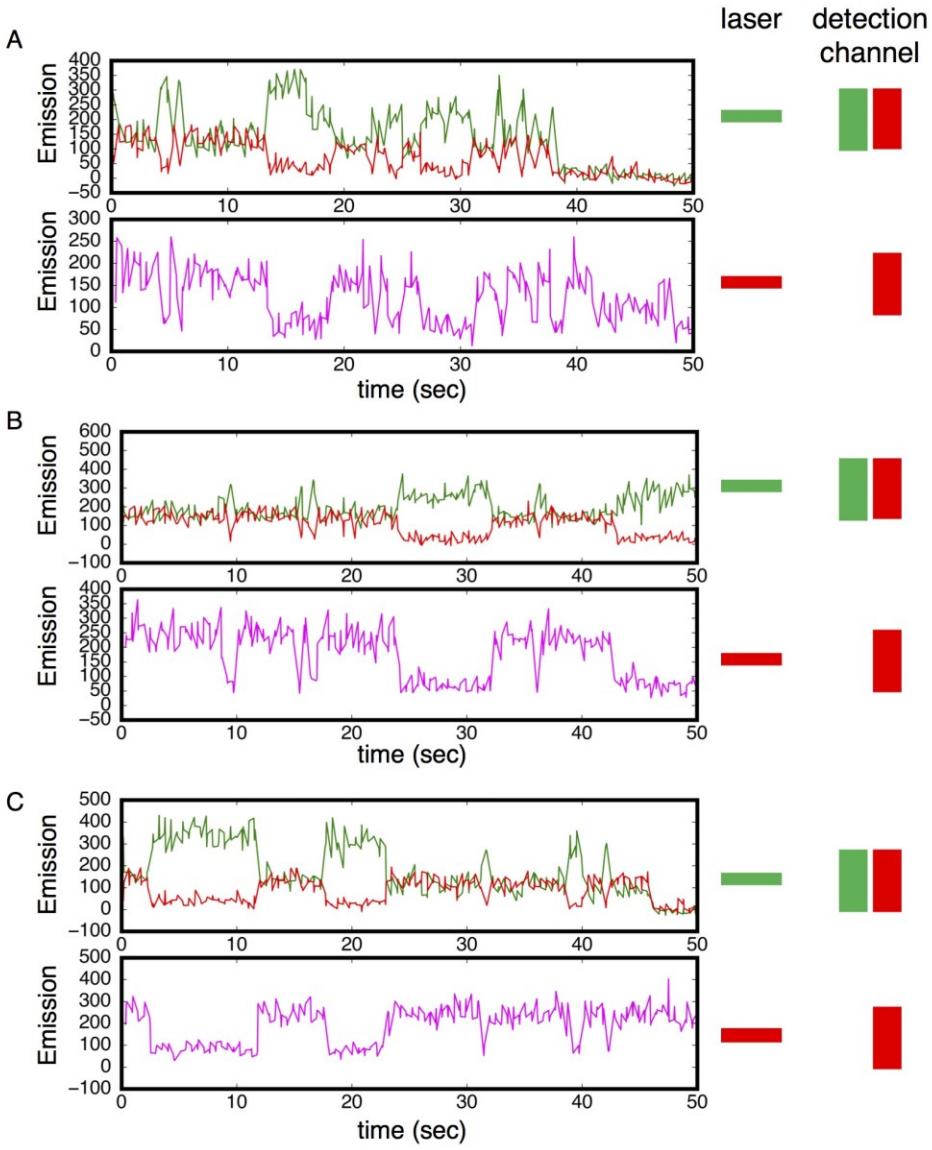


Figure 4.S1. cTALEs do not slide onto ends of short dsDNA. Apparent association and dissociation rate constants as a function of DNA concentration for an 8 repeat cTALE array binding to uncapped (blue) and capped (red) DNA, measured by single molecule FRET dwell time analysis. (A) Rate constants for conversion from the low to the high FRET state ($FRET_{L\rightarrow H}$). Rate constants for the faster phase are DNA concentration dependent (circles), indicating a bimolecular event. The DNA concentration-dependence is stronger (larger slope) with uncapped DNA, which may be due to faster diffusion of small, uncapped DNA (10 kDa) compared to large, capped DNA (320 kDa). Rate constants for the slower phase are not DNA-concentration dependent (triangles). (B) Rate constants for conversion from the high to the low FRET state ($FRET_{H\rightarrow L}$). Neither phase shows a DNA concentration dependent, indicating unimolecular steps. Conditions: 20 mM Tris pH 8.0, 200 mM KCl.

Figure 4.S2. Alternating laser experiments show agreement between cTALE$_8$ FRET and co-localization kinetics. (A-C) Three representative time trajectories showing similar FRET and co-localization profiles. Long movies collected alternating 5 frames green laser excitation (green and red channel emission shown in green and red respectively) followed by 5 frames red laser excitation (red channel emission shown in in pink) with 50 msec exposure time. After data collection, emission from green and red laser excitation were separated to generate plots (A-C). Conditions: 20 mM Tris pH 8.0, 200 mM KCl.


laser
detection channel
A
Emission
time (sec)
B
Emission
time (sec)
C
Emission
time (sec)

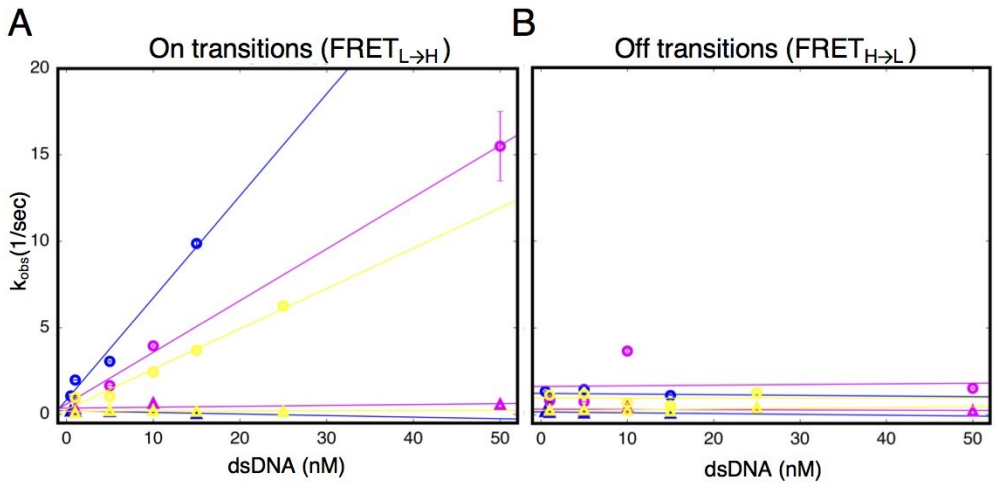
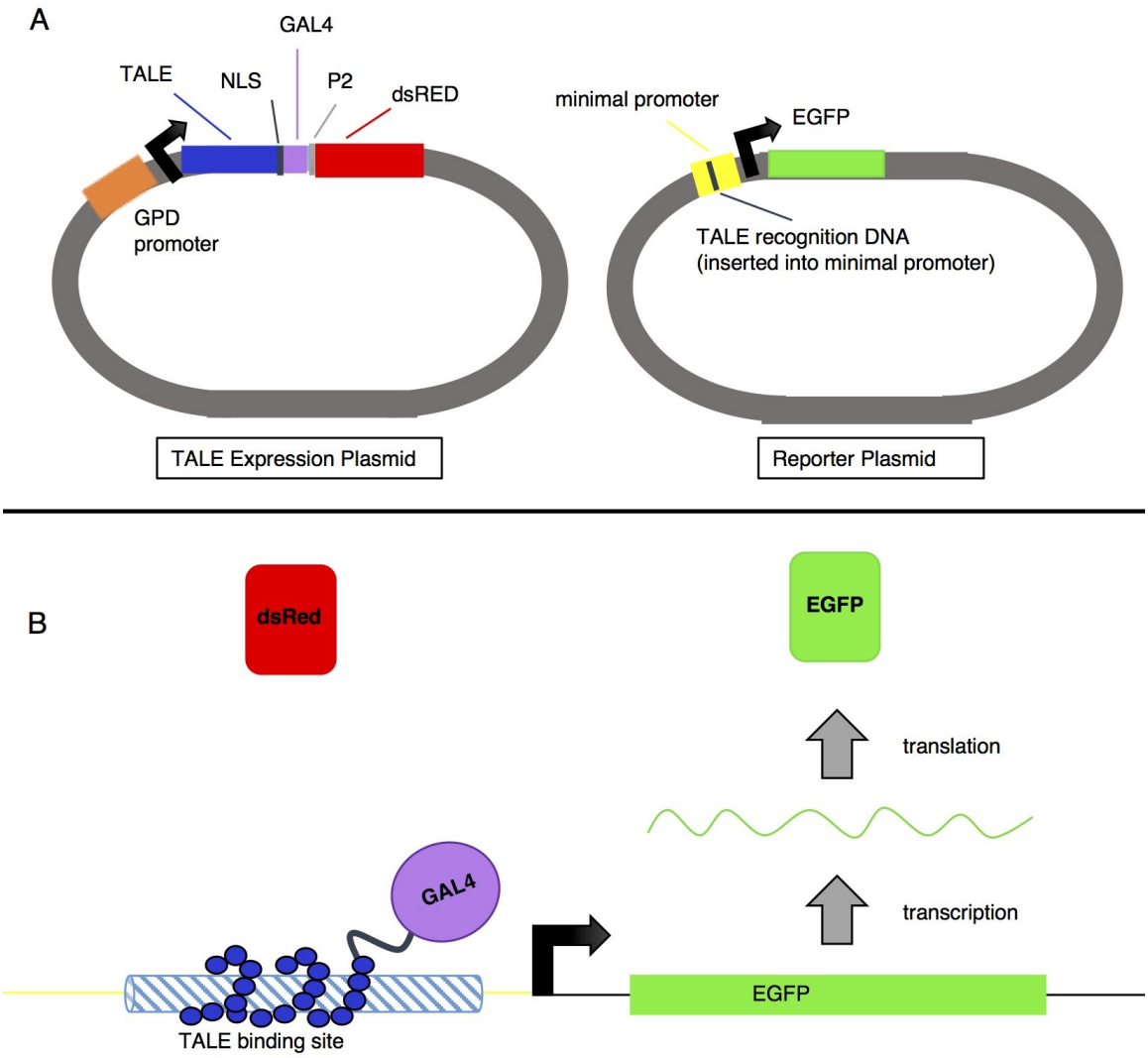


Figure 4.S3. Urea and destabilizing mutations decrease apparent binding rate of cTALE$_8$. (A) Apparent association rate constants as a function of DNA concentration for an 8 repeat cTALE in 0 M urea (blue), 1 M urea (pink), and with destabilizing point mutations (yellow). The DNA concentration-dependence is strongest (steepest slope) in the absence of urea and destabilizing point substitutions (circles). Rates in the slower phase (triangles) appear unaffected by urea or mutational destabilization. (B) Apparent dissociation rate constants as a function of DNA concentration (phase 1 shown in circles, and phase 2 shown in triangles). Rates in the both phases appear unaffected by urea and mutational destabilization. Conditions: 20 mM Tris pH 8.0, 200 mM KCl.

Figure 4.S4. Schematic of *S. cerevisiae* reporter plasmids and assay. (A) TALE expression plasmid and reporter plasmid used to measure transctiption activation. The TALE array is expressed under control of GDP promoter to drive high protein expression. The TALE array is fused to an nuclear localization sequences (NLS), a GAL4 transcription activation domain (GAL4), and dsRED, which is separated by P2 peptide. Cleavage of the P2 site generates two separate proteins: TALE~GAL4 and dsRed (used as a proxy for TALE protein production). The reporter plasmid contains a minimal promoter[26] including to ten cognate TALE binding sites controlling expression of EGFP. (B) The TALE protein binds cognate sites adjacent to the synthetic minimal yeast promoter and activates transcription of reporter EGFP. In the absence of TALE-DNA binding, transcription of EGFP is not activated. Quantification of EGFP (reporter) and dsRED (transformation control) mRNA is performed using qPCR.


A
GAL4
TALE
NLS
P2
dsRED
GPD
promoter
TALE Expression Plasmid
minimal promoter
EGFP
TALE recognition DNA
(inserted into minimal promoter)
Reporter Plasmid
B
dsRed
EGFP
translation
transcription
GAL4
EGFP
TALE binding site

CONCLUSION

Nature tunes folding cooperativity for function. Proteins must be stable and cooperative enough to prevent disease-causing aggregation, but this must be balanced at times to generate populations of partly folded states required for function. Unnatural helical repeat proteins called *De novo* Helical Repeats (DHRs), where the sequence and structure are unlike any natural-occurring proteins, fold cooperatively. However, unlike naturally-derived repeat proteins, DHRs partition global stability into both intrinsic and interfacial energies. As a result, DHRs are extremely stable and superfast folding.

Cooperativity in naturally-occurring repeat protein, transcription activator-like effectors (TALEs), is tuned for function. TALEs are bacterial virulence factors injected into plants. A repeat domain is responsible for sequence-specific DNA recognition resulting in transcriptional activation of plant genes vital for bacterial survival. TALEs wrap superhelically around DNA binding one repeat per base pair. Specificity of each repeat is controlled by two loop residues (position 12 and 13 called RVDs) making sequence specific contacts with the major groove. Surprisingly, changes to RVD sequence affects both the stability and cooperativity of TALE arrays. TALE proteins populate several types of partly folded states.

Population of partly folded TALE protein states affects DNA binding kinetics. DNA-binding kinetics confirm conformational heterogeneity in free and DNA-bound TALE proteins. TALE proteins containing more than 11 repeats form more than one superhelical turn around DNA, and bind DNA via a high energy

“open” state. Most TALEs do form multiple superhelical turns, given the average number of repeats in a TALE gene is 17.5. Conformational heterogeneity in the DNA-bound state is present in long and short TALEs, and may represent a specificity checking mechanism. Future work will elucidate structural models for conformational heterogeneity in TALE DNA binding kinetics.

VITA

Kathryn Rachelle Geiger-Schuller was born on April 4, 1988 in Indianapolis, Indiana. She attended Amy Beverland Elementary School and Belzer Middle School. She graduated from Lawrence Central High School in 2007. She attended Indiana University in Bloomington, Indiana and worked in the laboratory of Dr. David Giedroc. She graduated in 2011 with a Bachelor of Arts in French and a Bachelor of Science in Biochemistry. In the fall of 2011 she joined the Program in Molecular Biophysics at The Johns Hopkins University in Baltimore, Maryland to pursue her Ph.D. She completed her dissertation research in the laboratory of Dr. Doug Barrick.